IN4325

Neural IR

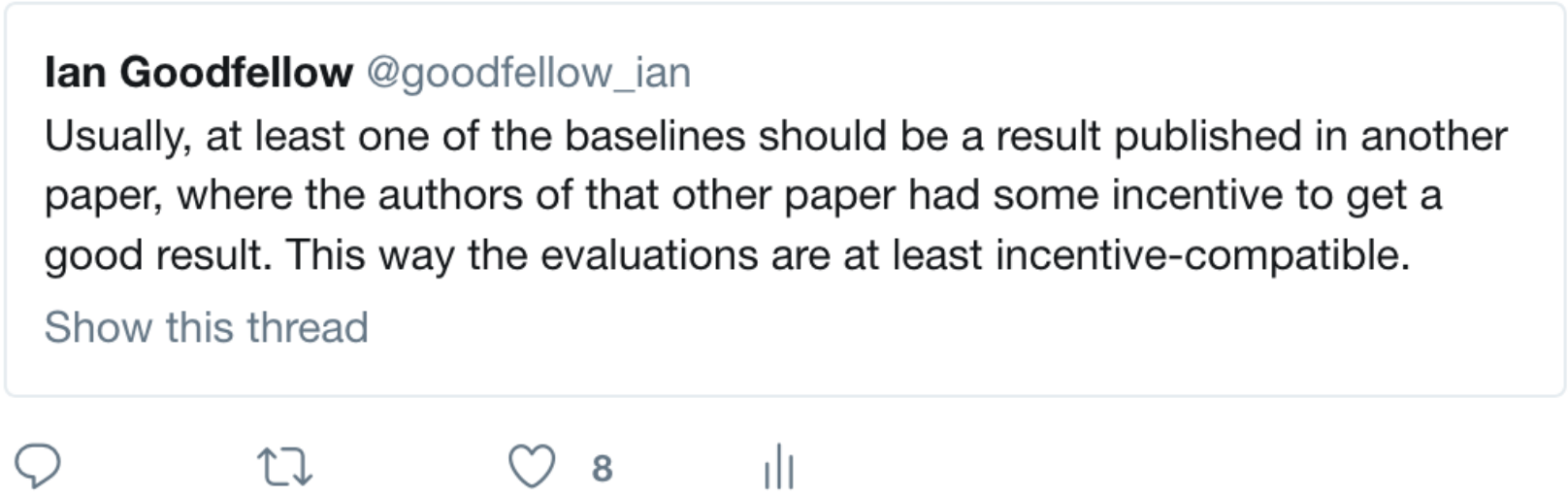
Arthur Câmara (WIS, TU Delft)
Slides (mostly) by Claudia Hauff

ECIR 2019 Cologne

39 Full Papers. (23% acceptance rate)

- 3 contain “neural” in title
- 3 contain “embedding” in title
- 11 others **_may_** contain neural “stuff”, given title.

~43% of papers with some type of neural IR.



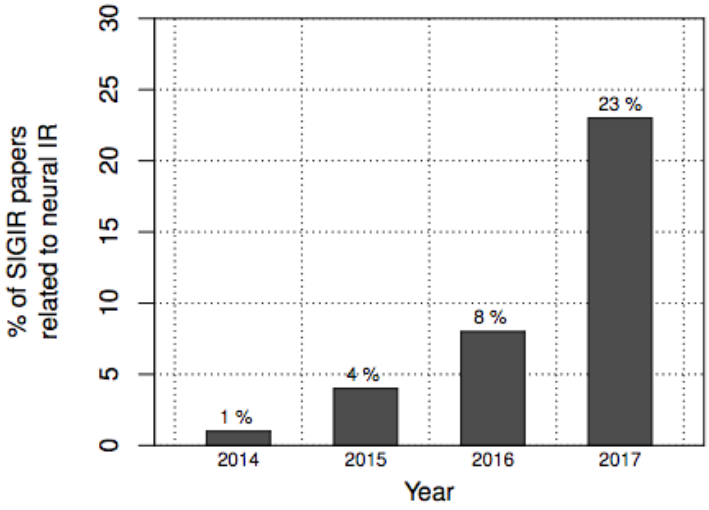
Different fields have different standards ...

https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

A great starting point for Neural IR - 123 pages of insights! We follow some of it here.

Neural IR

… is not taking over the IR research field **(yet)**



Covered on a high level: this topic by itself could take up all whole quarter of lectures. Unfortunately, we do not have the time.

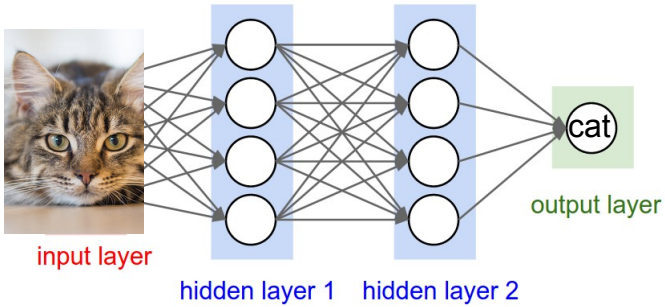
http://cs231n.github.io/neural-networks-1/

https://arxiv.org/pdf/1409.4842.pdf

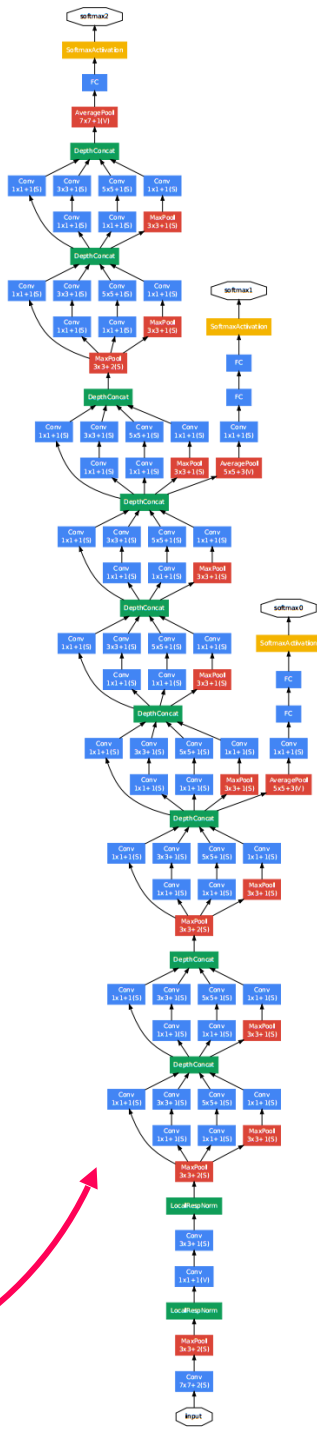
Overview

`NeuralIR` is the application of shallow or deep neural networks to IR tasks.

`NeuralIR` can make use of neural NLP techniques to enhance the retrieval Pipeline.



GoogLeNet (2014)



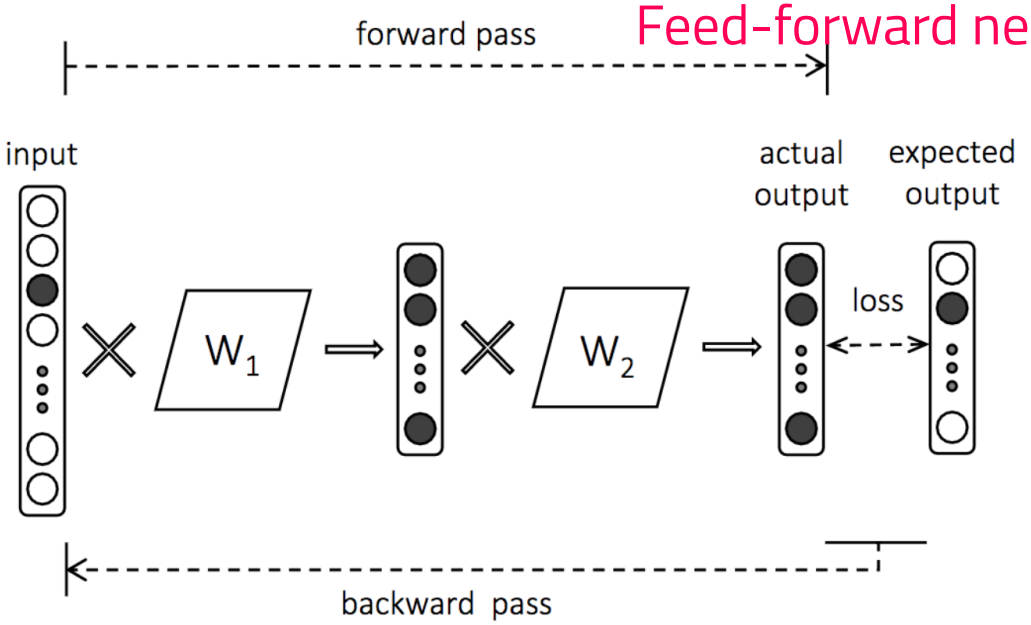
#params
LM, BM25

`NeuralIR` models contain thousands/millions of parameters. They require a large amount of training data.

Instead of hand-crafting **features** (classic ML), we now handcraft NN **architectures**

https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

Neural net basics



Feed-forward neural net with one hidden layer.

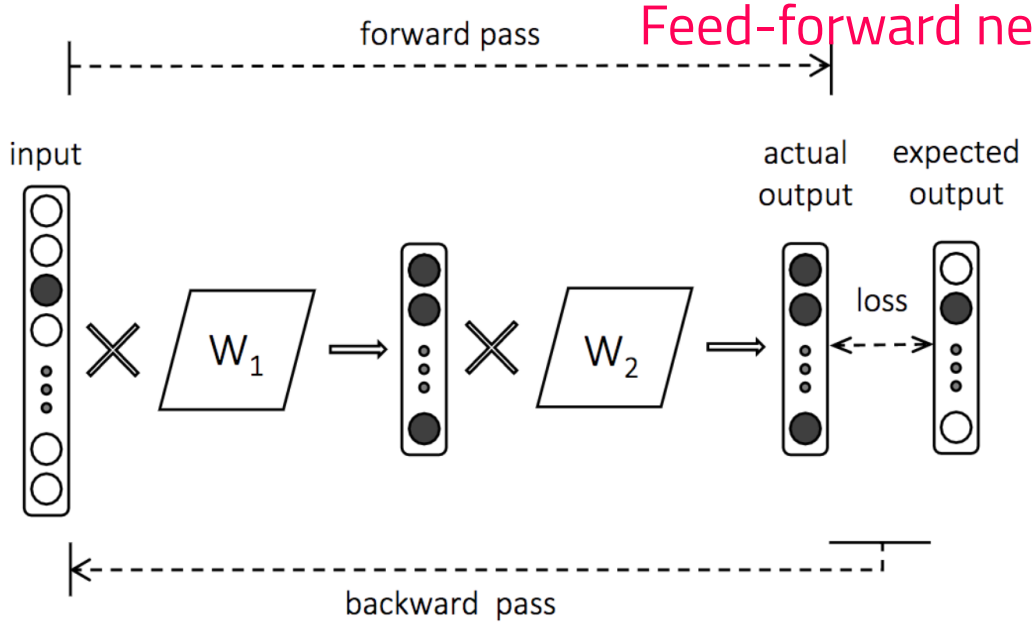
https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

Initially: deep nets save us from expensive and elaborate feature engineering.

Now: lets engineer architectures and hardware/software that allows us to efficiently train deep nets.

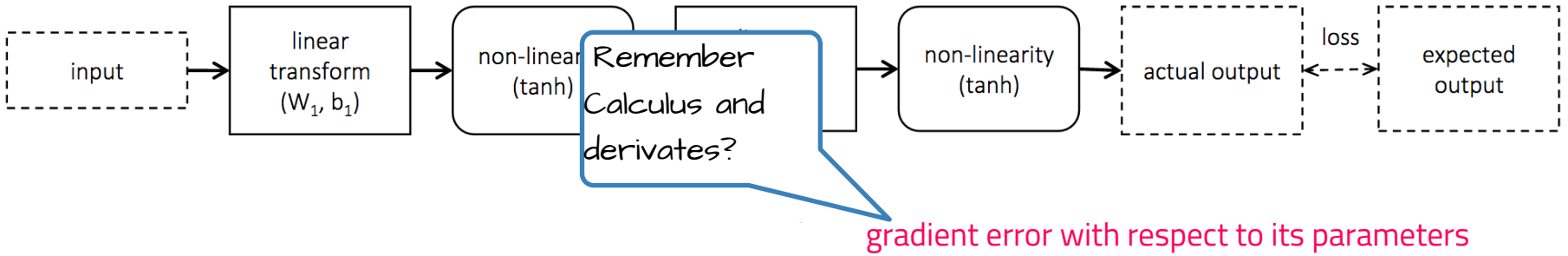
Neural net basics

Feed-forward neural net with one hidden layer.



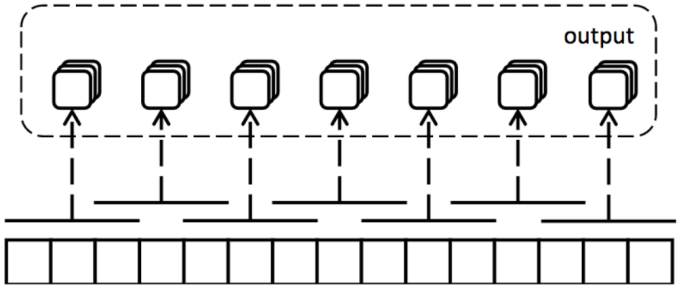
$$\vec{y} = tanh(W_2 \cdot tanh(W_1 \cdot \vec{x} + \vec{b}_1) + \vec{b}_2)$$

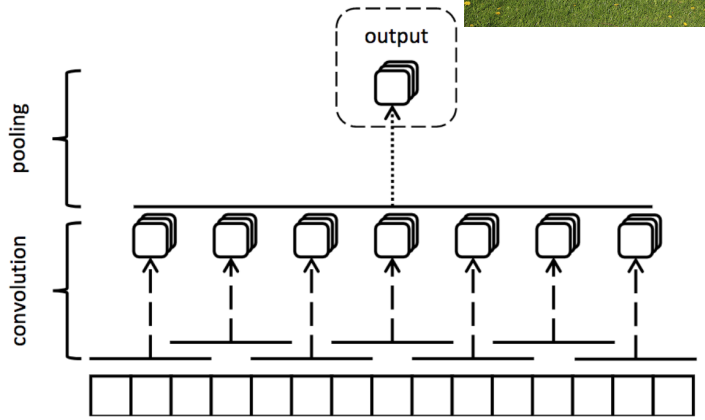
Chain of computational steps.



https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

CNNs and RNNs

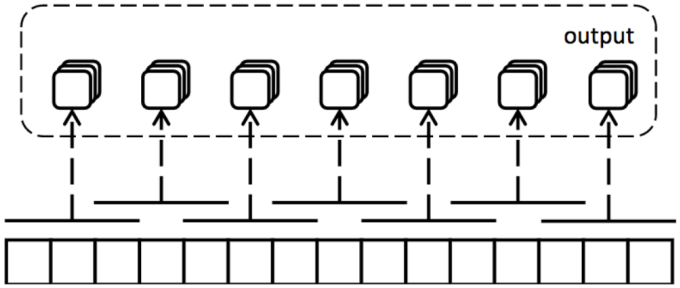


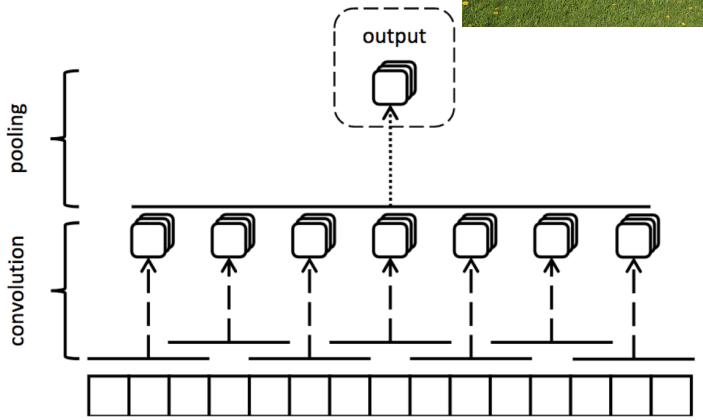


CNN: convolutional neural network.
Most often found in computer vision setups.

https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

CNNs and RNNs

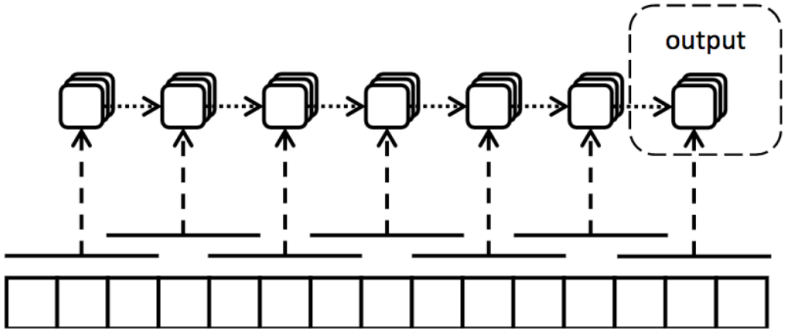




CNN: convolutional neural network.
Most often found in computer vision setups.

* Maybe should be revisited? (ICLR'18): https://openreview.net/pdf?id=BJEX-H1Pf

RNN: recurrent neural network.
Most often found in NLP setups.*



Obama entered the *White House*. *He* met *there* with Trump.

...ons, including convolutional and recurrent layers move a fixed size window over the input space with fixed stride.

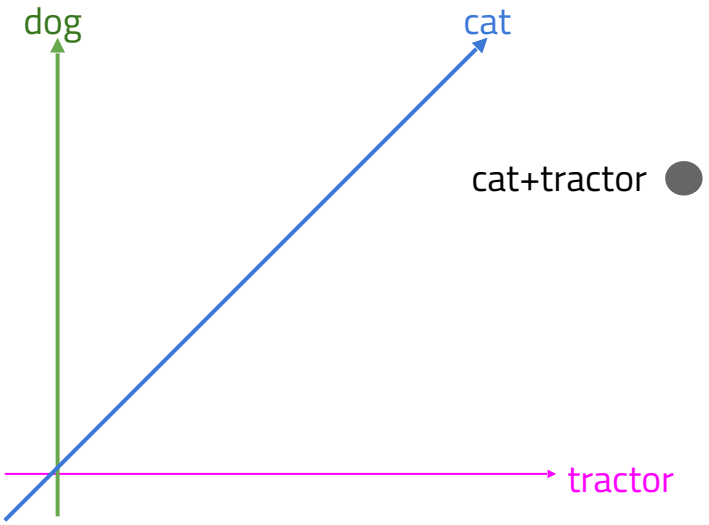
Each window is projected by a parameterized neural operation, often followed by an aggregation step such as (max) pooling.

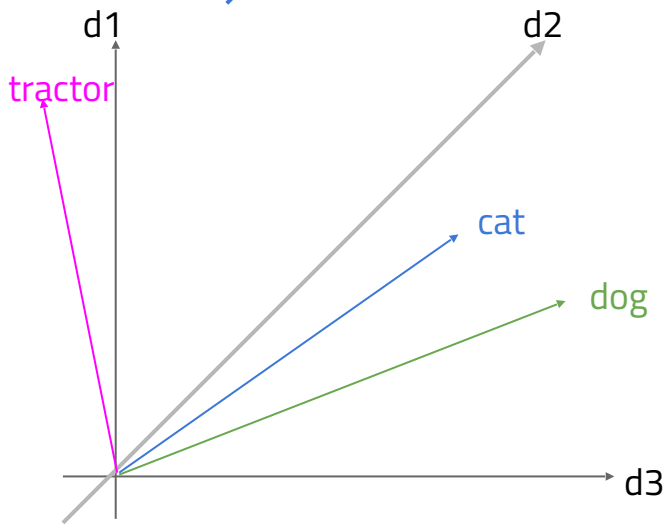
Text representations

Neural IR

https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

Text representations





In DL, commonly known as "one-hot" encoding/repr.

Local vector representation (sparse, high-dimensional)

$$cat = (0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ..., 0)$$
$$tractor = (0, 0, 0, 1, 0, 0, 0, 0, 0, 0, ..., 0)$$
$$dog = (1, ..., 0)$$
$$feline = (0, ..., 0)$$

Individual dimensions are atomic and interpretable.

Semantic gap: the similarity between any of

Individual dimensions are (usually) not interpretable

distributed vector representation (dense, often low-dimensional)

$$cat = (0, 1, 1, 0, 1, 0, 0, 0, 0, 0)$$
$$tractor = (0, 0, 0, 1, 0, 0, 0, 1, 1, 0)$$
$$dog = (1, 1, 0, 0, 0, 0, 0, 0, 0, 0)$$
$$feline = (0, 1, 1, 1, 1, 0, 0, 0, 0, 1)$$

The similarity between some of those terms is no longer zero.

https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

Text representations

In DL, commonly known as "one-hot" encoding/repr.

Local vector representation (sparse, high-dimensional)

$$cat = (0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ..., 0)$$
$$tractor = (0, 0, 0, 1, 0, 0, 0, 0, 0, 0, ..., 0)$$
$$dog = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ..., 0)$$
$$feline = (0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ..., 0)$$

Individual dimensions are atomic and interpretable.

Text representations can be learnt in a **supervised** or **unsupervised** fashion.

In IR, supervised approaches use query-document pairs.

In IR, the unsupervised approach uses just queries or just documents.

"**Similarity**" is not an absolute concept, it depends on the task & context at hand. In IR, it is related to relevance.

Semantic gap: the similarity between any of

Individual dimensions are (usually) not interpretable

distributed vector representation (dense, often low-dimensional)

$$cat = (0, 1, 1, 0, 1, 0, 0, 0, 0, 0)$$
$$tractor = (0, 0, 0, 1, 0, 0, 0, 1, 1, 0)$$
$$dog = (1, 1, 0, 0, 0, 0, 0, 0, 0, 0)$$
$$feline = (0, 1, 1, 1, 1, 0, 0, 0, 0, 1)$$

The similarity between some of those terms is

Extreme case: very long document with a single relevant sentence. Still relevant ...

https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

Feature-based representation examples

Which representation is best so that:
- "Amsterdam" and "London"
- "Ajax" and "Amsterdam"

are similar? (**Typical vs Topical**!)

in-document features

$$cat = (0, 0, \underset{\downarrow}{1}_{doc\ 3}, 0, \underset{\downarrow}{1}_{doc\ 5}, 0, 0, 0, 1, 0, ..., \underset{\downarrow}{1}_{doc\ 1678}, 0)$$

neighbouring-word features

$$cat = (0, 0, \underset{\downarrow}{1}_{runs}, 0, \underset{\downarrow}{1}_{eats}, 0, 0, 0, \underset{\downarrow}{1}_{hurt}, 0, ..., \underset{\downarrow}{1}_{is}, 0)$$

neighbouring-word with distances features

$$cat = (0, 0, \underset{\downarrow}{1}_{runs^{+1}}, 0, \underset{\downarrow}{1}_{black^{-1}}, 0, 0, 0, \underset{\downarrow}{1}_{along^{+2}}, 0, ..., \underset{\downarrow}{1}_{old^{-1}}, 0)$$

character trigram features

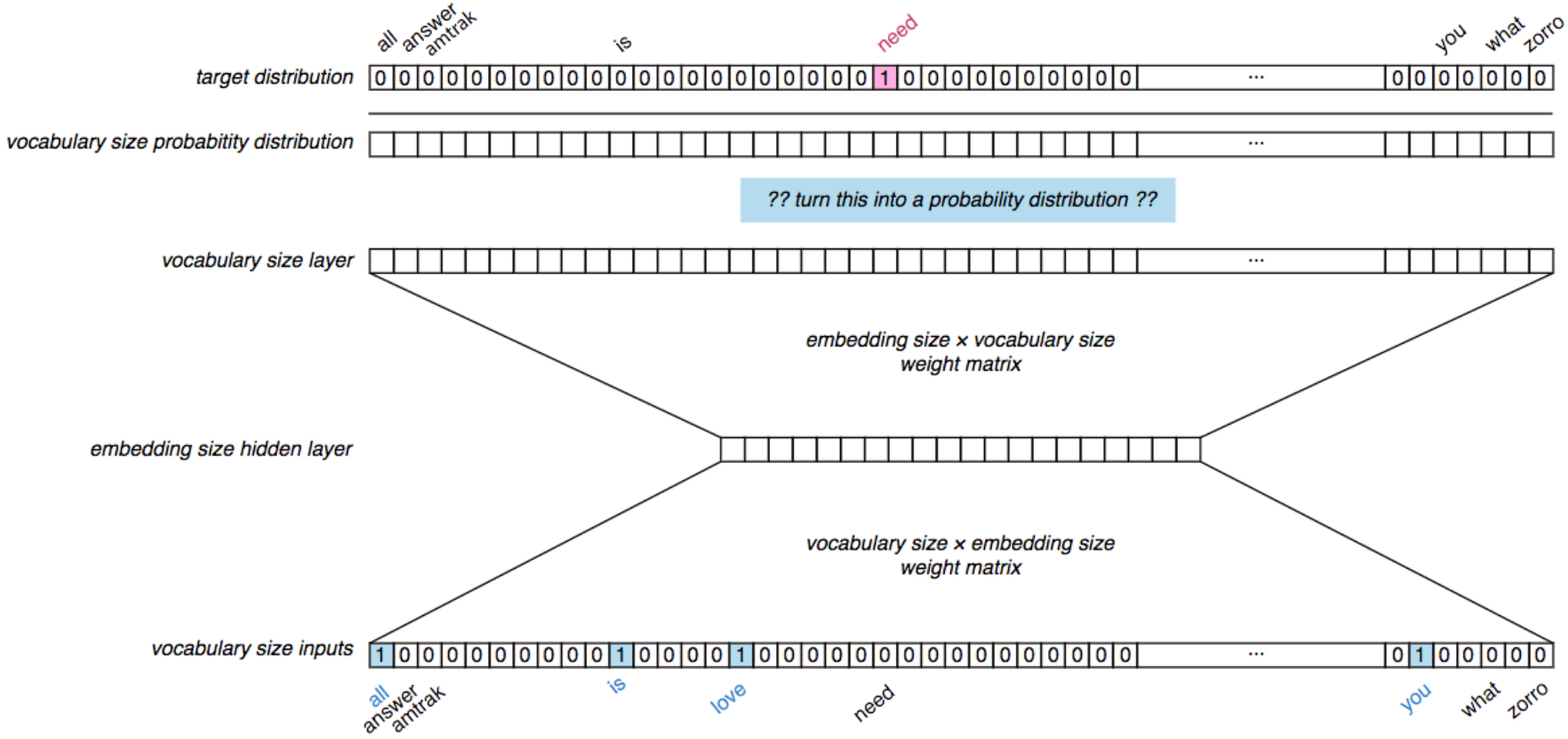
$$kitten = (0, 0, \underset{\downarrow}{1}_{en\#}, 0, \underset{\downarrow}{1}_{\#ki}, 0, \underset{\downarrow}{1}_{tte}, 0, \underset{\downarrow}{1}_{kit}, 0, \underset{\downarrow}{1}_{ten}, ..., \underset{\downarrow}{1}_{itt}, 0)$$

http://nn4ir.com/ecir2018/slides/02_Preliminaries.pdf (image source)

Embeddings

word2vec/GloVe Will be covered in more details by Nava!

word2vec/GloVe: the vector of a word should be similar to the vectors of its neighbouring words



... representation in a new space that should preserve the relationships between items of the original representation.

Embeddings

Dense vector representation.

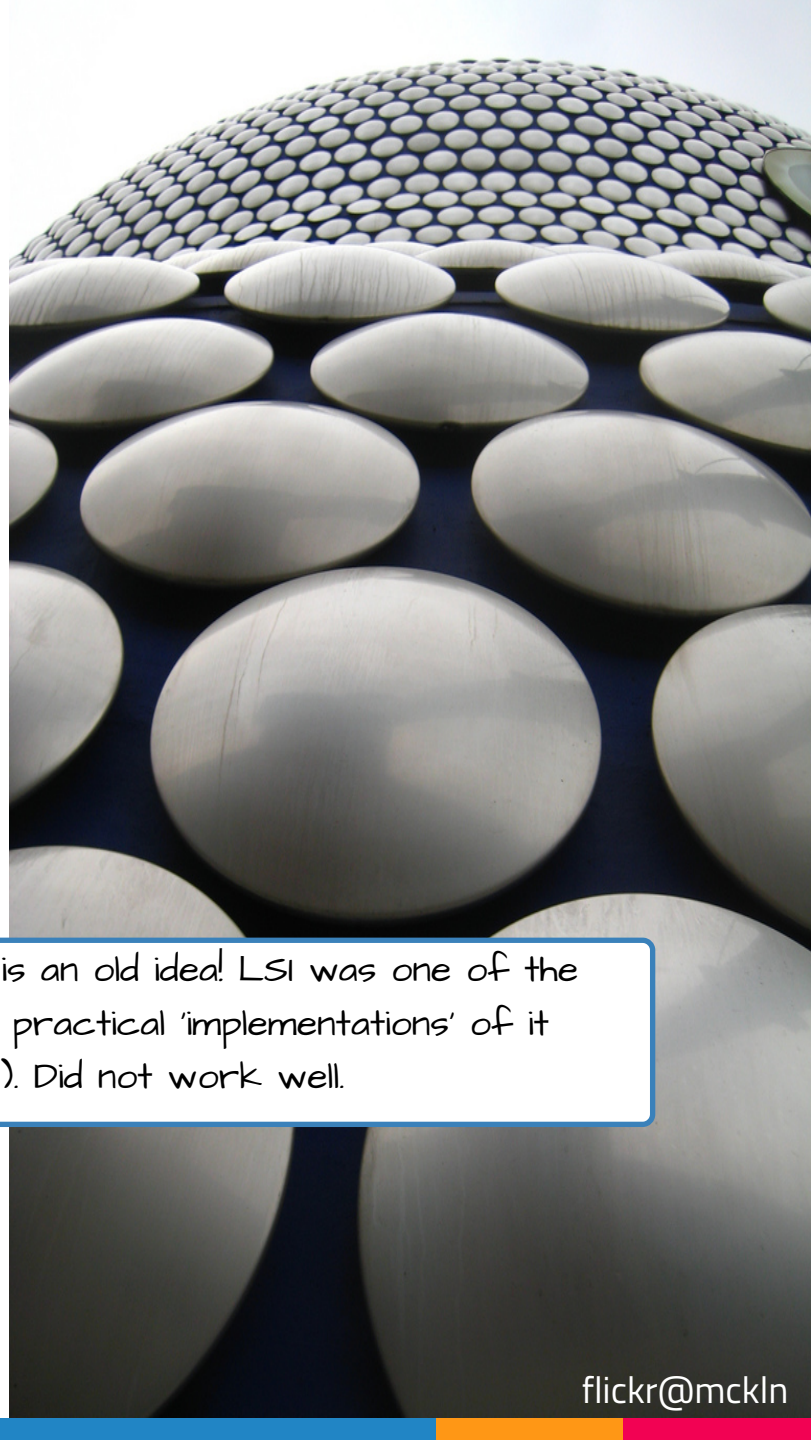
Low-dimensional.

Learnt from data.

Distributed representation most often refers to learnt embeddings.

In today's NLP literature, the default encoding.

Motivated by the **distributional hypothesis** (Harris, 1954): "*a word is characterized by the company it keeps*" (Firth, 1957)



This is an old idea! LSI was one of the first practical 'implementations' of it (1988). Did not work well.

https://arxiv.org/pdf/1610.08136.pdf

Local *and* distributed representations are needed

The President of the United States of America (POTUS) is the elected head of state and head of government of the United States. The president leads the executive branch of the federal government and is the commander in chief of the United States Armed Forces. Barack Hussein Obama II (born August 4, 1961) is an American politician who is the 44th and current President of the United States. He is the first African American to hold the office and the first president born outside the continental United States.

Q="united states president"

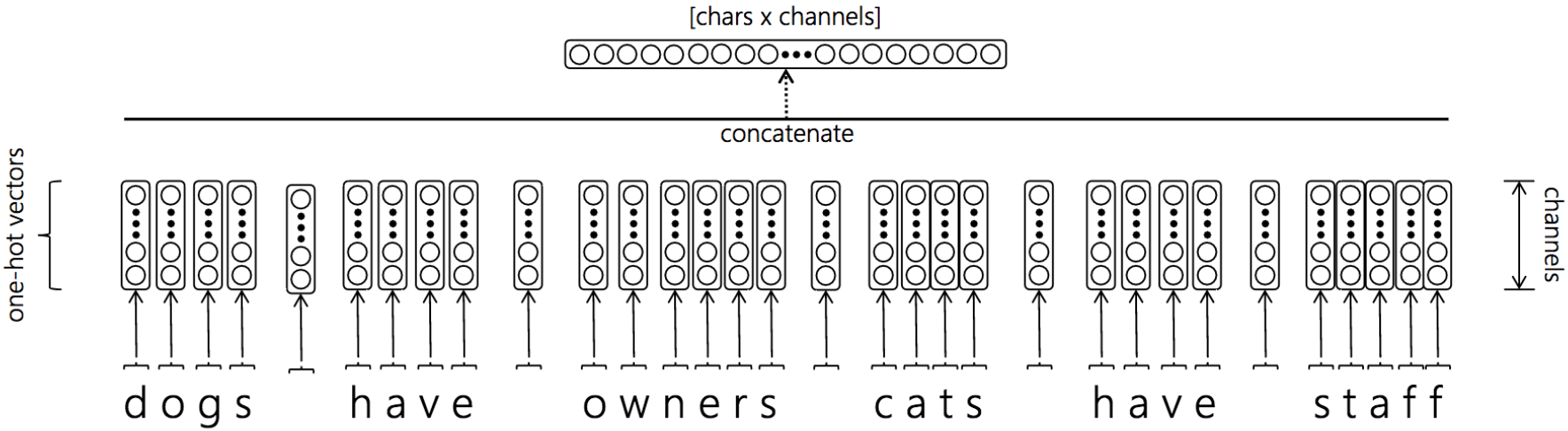
Shade of green indicates the drop in retrieval model's document score by individually removing each of the passage terms.

The President of the United States of America (POTUS) is the elected head of state and head of government of the United States. The president leads the executive branch of the federal government and is the commander in chief of the United States Armed Forces. Barack Hussein Obama II (born August 4, 1961) is an American politician who is the 44th and current President of the United States. He is the first African American to hold the office and the first president born outside the continental United States.

https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

Text input to deep neural nets

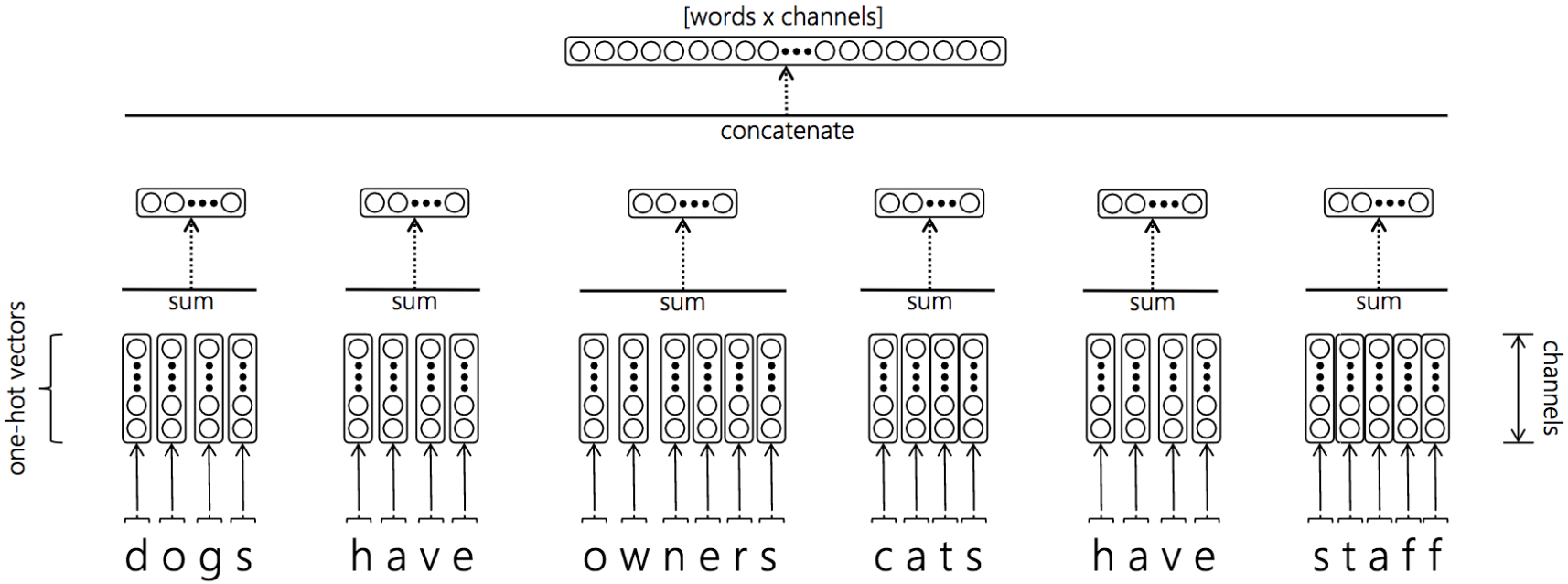
character-level input



https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

Text input to deep neural nets

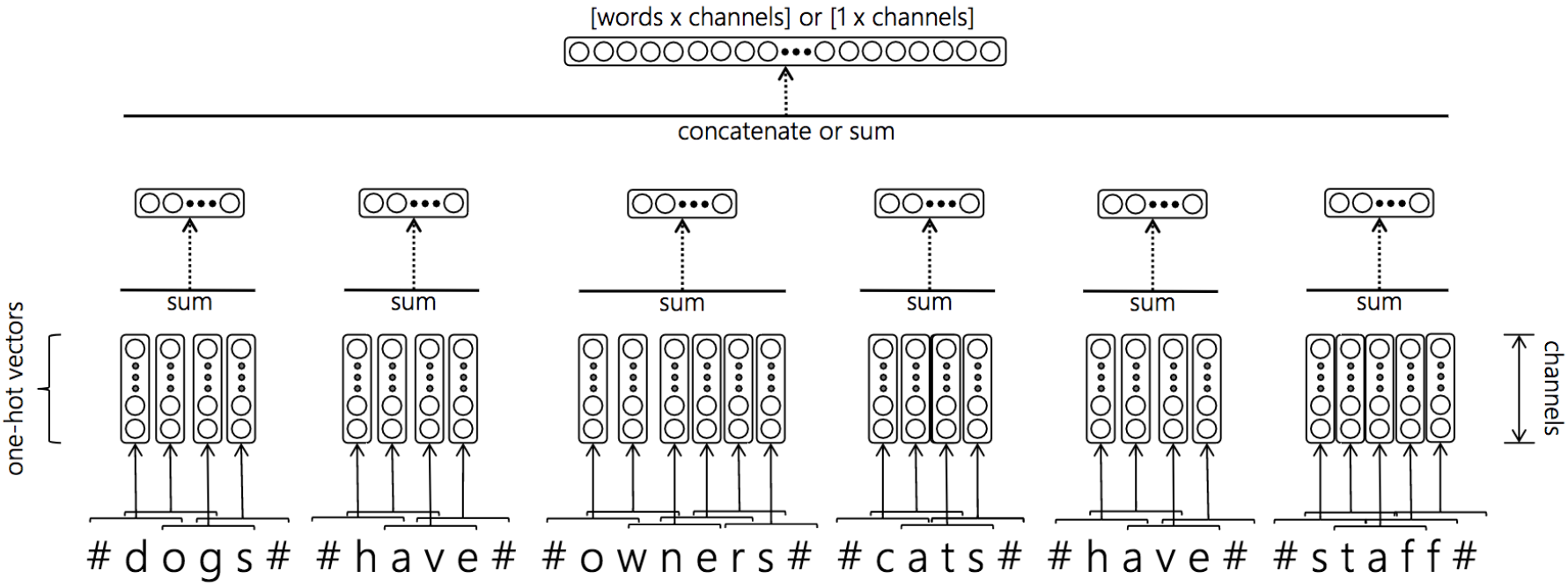
term-level input with bag-of-chars per term



https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

Text input to deep neural nets

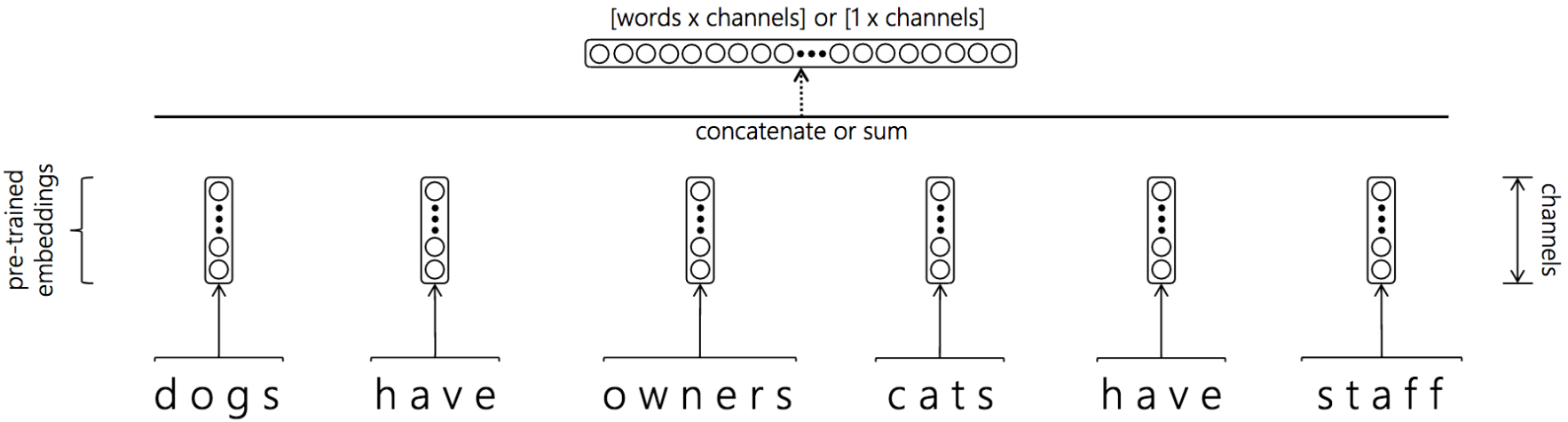
term-level input with bag-of-trigrams per term



https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

Text input to deep neural nets

term-level input with pre-trained word embeddings



`GloVe` (Nava will explain how this works exactly)

- Pre-trained word vectors. This data is made available under the Public Domain Dedication and License v1.0 whose full text can be found at: http://www.opendatacommons.org/licenses/pddl/1.0/.
  - Wikipedia 2014 + Gigaword 5 (6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors, 822 MB download): glove.6B.zip
  - Common Crawl (42B tokens, 1.9M vocab, uncased, 300d vectors, 1.75 GB download): glove.42B.300d.zip
  - Common Crawl (840B tokens, 2.2M vocab, cased, 300d vectors, 2.03 GB download): glove.840B.300d.zip
  - Twitter (2B tweets, 27B tokens, 1.2M vocab, uncased, 25d, 50d, 100d, & 200d vectors, 1.42 GB download): glove.twitter.27B.zip

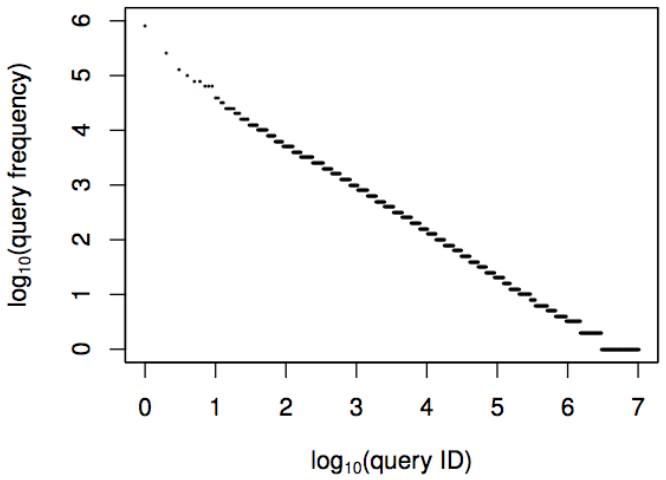
https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

Neural IR architectures

https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

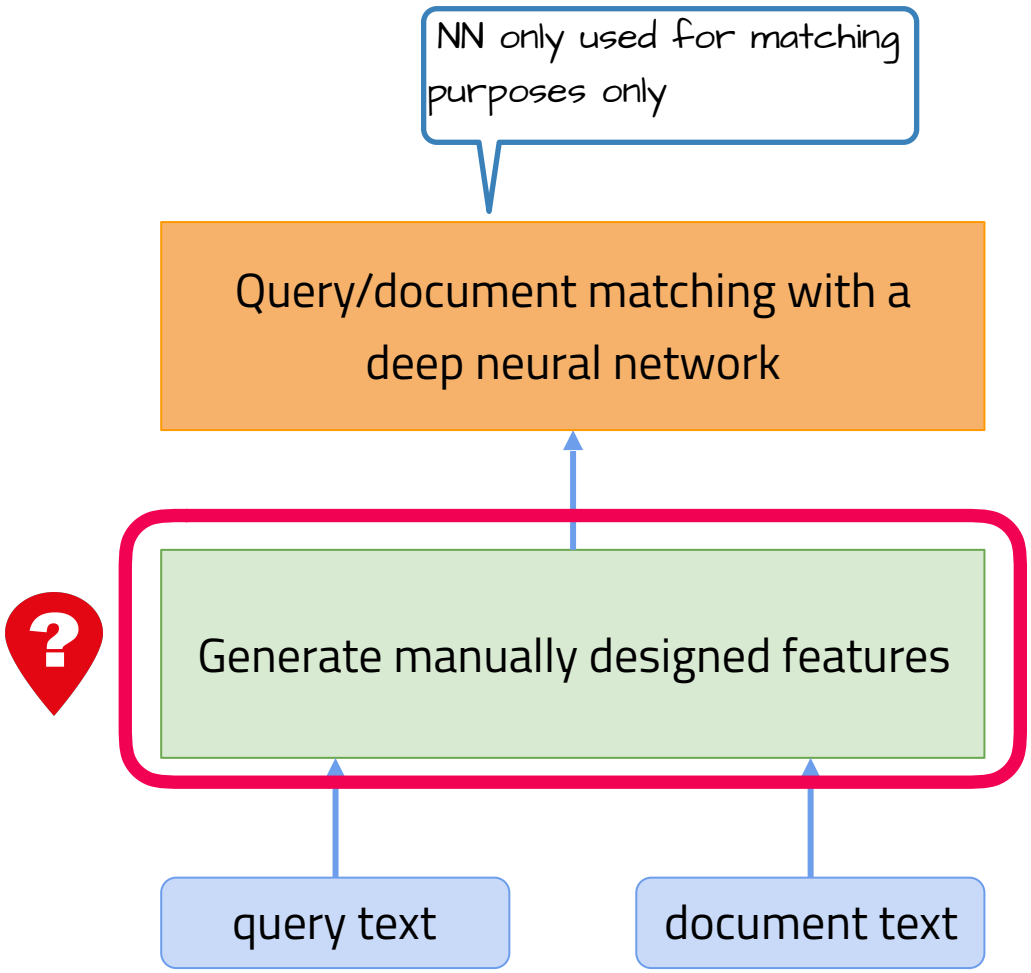
A good retrieval system should …

1. Have **semantic understanding** and enable **exact term matching** (*cat* and *feline* are similar query terms, but *hot dog* is not similar to *warm puppy*; *[rare name] gardening*)
2. Be **robust** to rare **inputs**: remember the long tail
3. Be **robust** to **corpus variance**: BM25 out-of-the-box is a good ranking model for all domains, neural models require a lot of training (training on domain X data and testing on domain Y data can go wrong ….)



4. Be **robust** to **variable length input** (classic IR has document length normalization)
5. Be **sensitive** to **context** (e.g. location)

Learning to rank



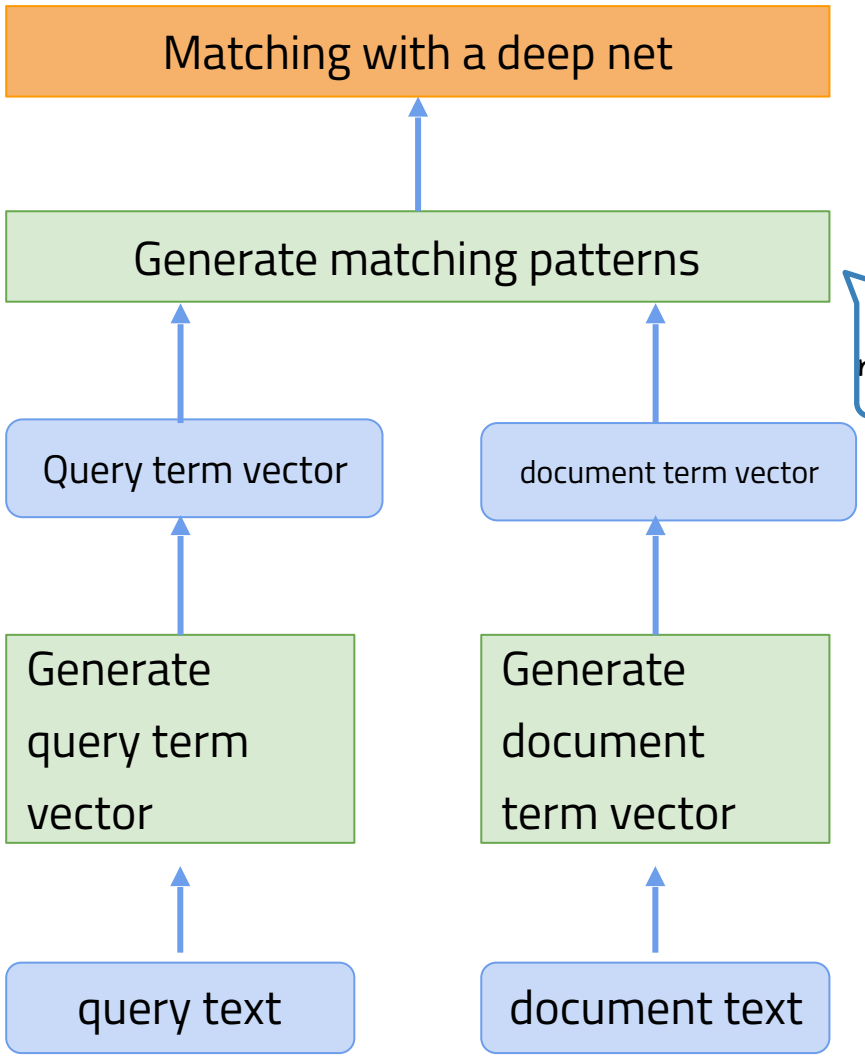
Categorized according to their:

- Query representation
- Document representation
- Relevance estimation

Interaction focused models

Categorized according to their:

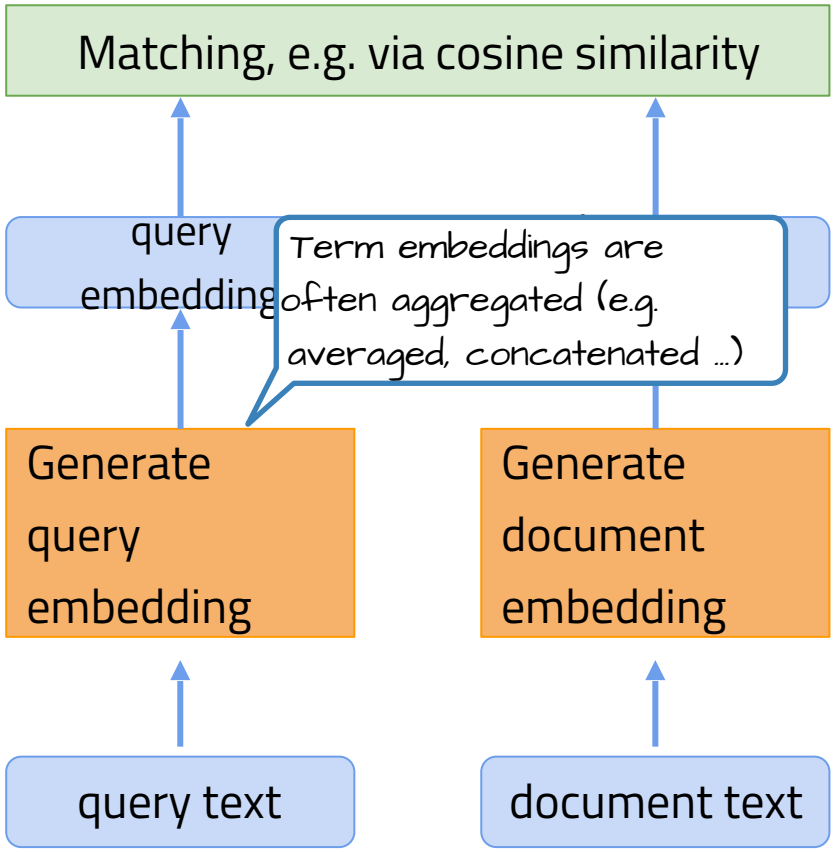
- Query representation
- Document representation
- Relevance estimation



Representation focused models

Categorized according to their:

- Query representation
- Document representation
- Relevance estimation



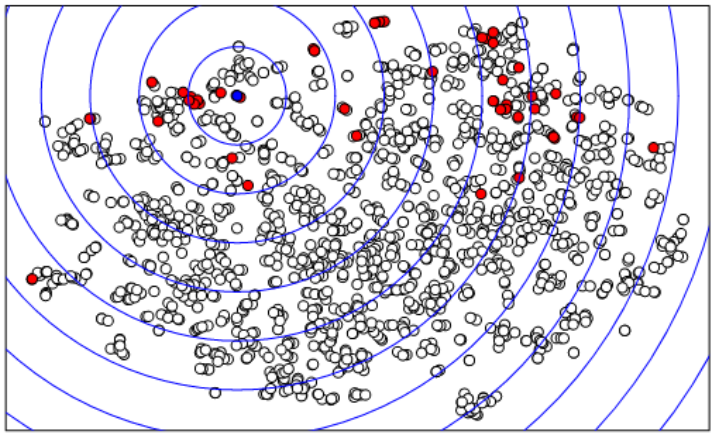
$$sim(q,d) = cos(\vec{v}_q, \vec{v}_d) = \frac{\vec{v}_q^{\intercal} \vec{v}_d}{\|\vec{v}_q\| \|\vec{v}_d\|}$$

Query expansion

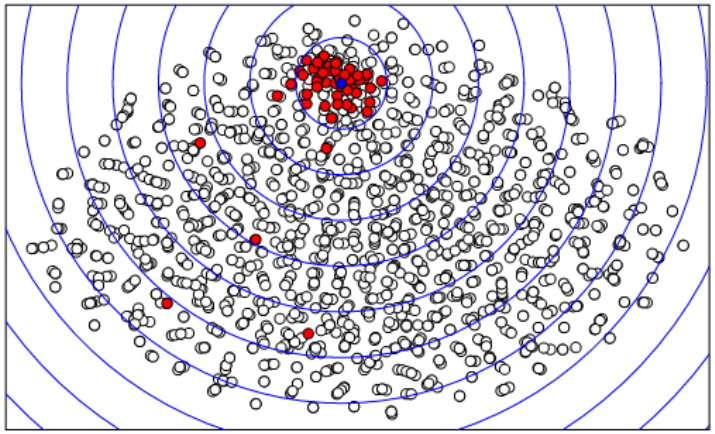
Categorized according to their:

- Query representation
- Document representation
- Relevance estimation

global



local



Point: term

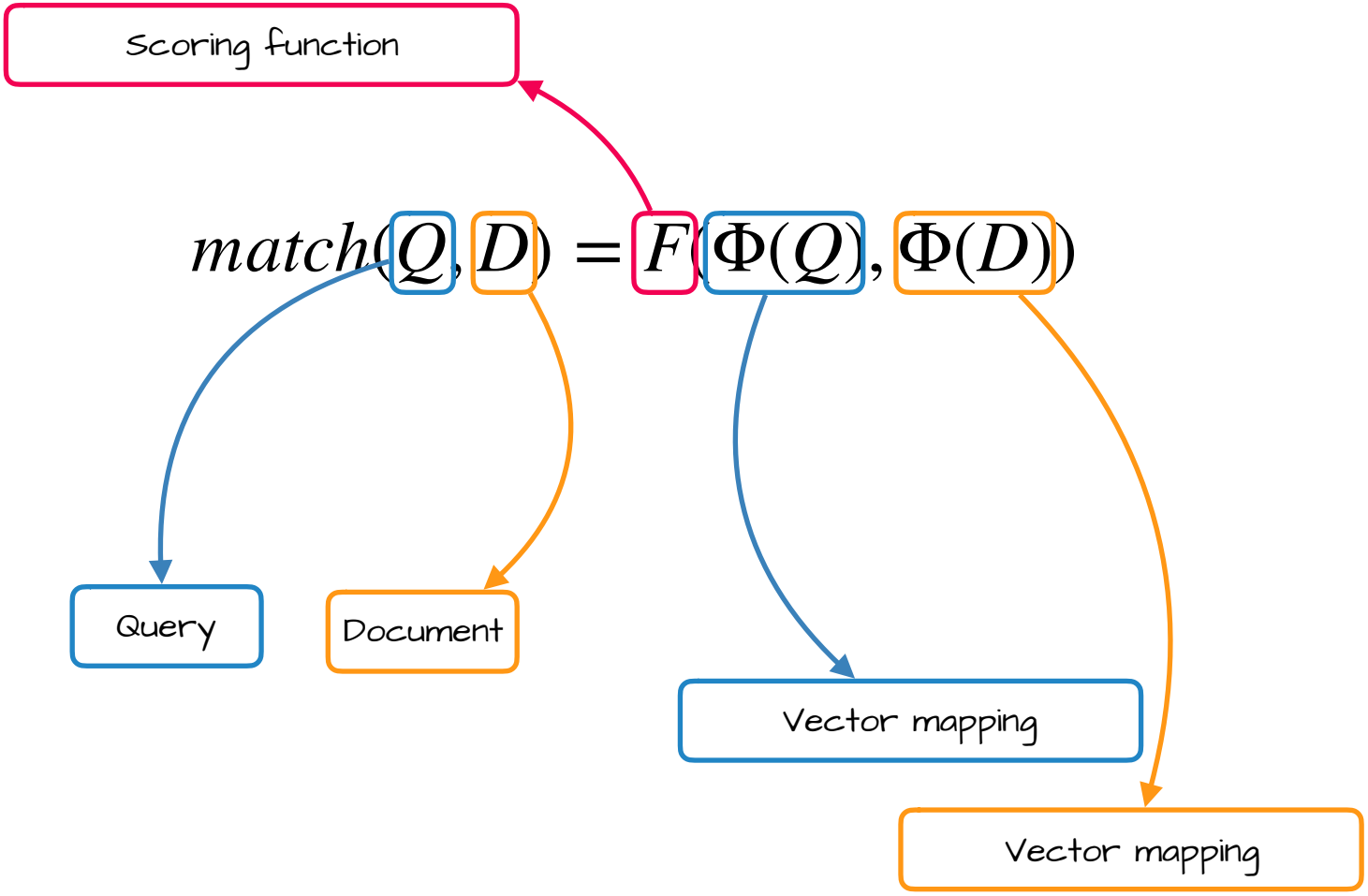
Blue: query term

Red: term occurring in the relevant set of documents

t-SNE projection (stochastic neighbour embedding)

A simple idea that works well for PRF: learn embeddings on the top retrieved documents only instead of a large corpus.

Generally speaking...



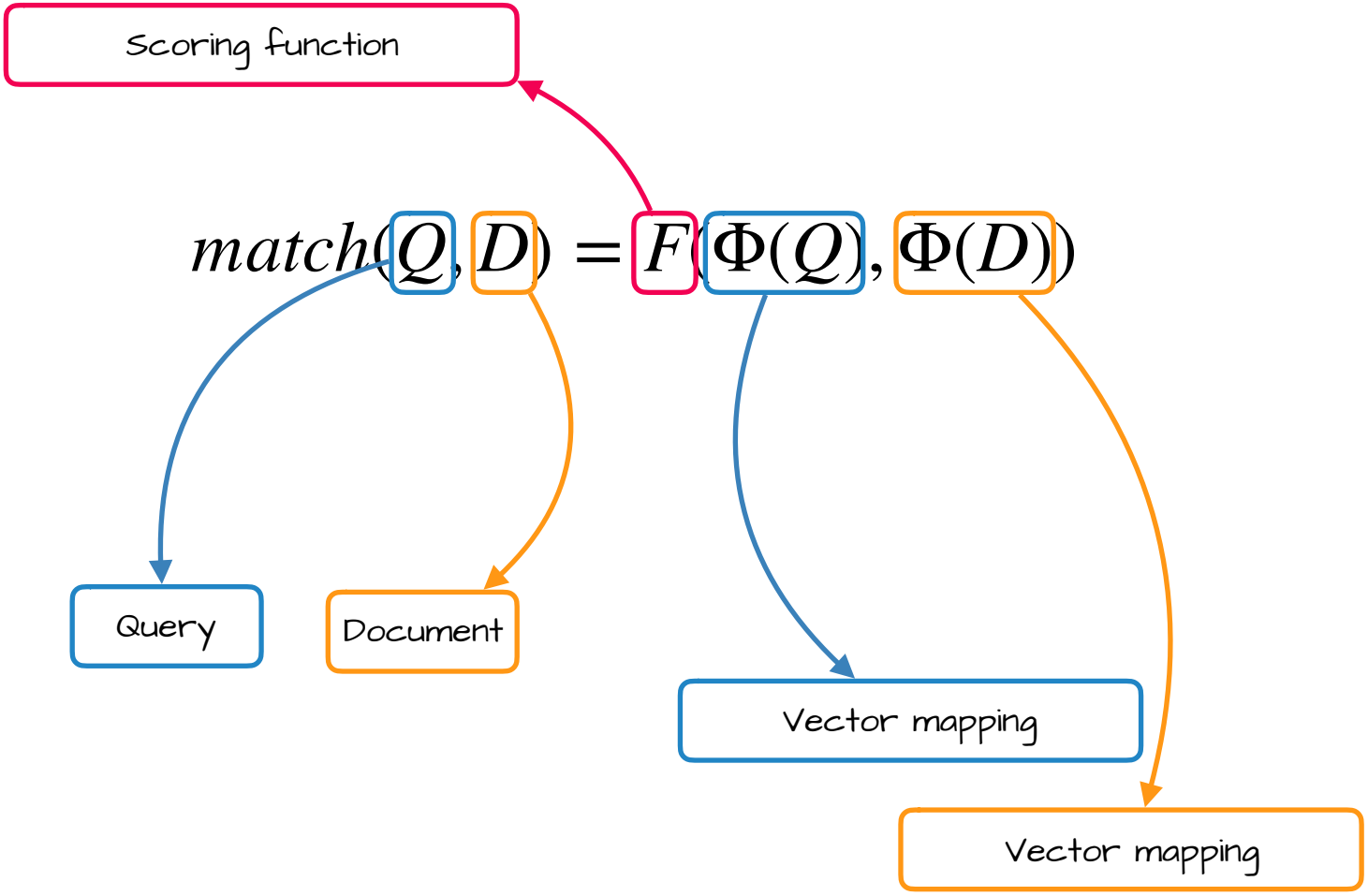
https://dl.acm.org/citation.cfm?id=2767830

Representation focused models

(Or Siamese Networks)

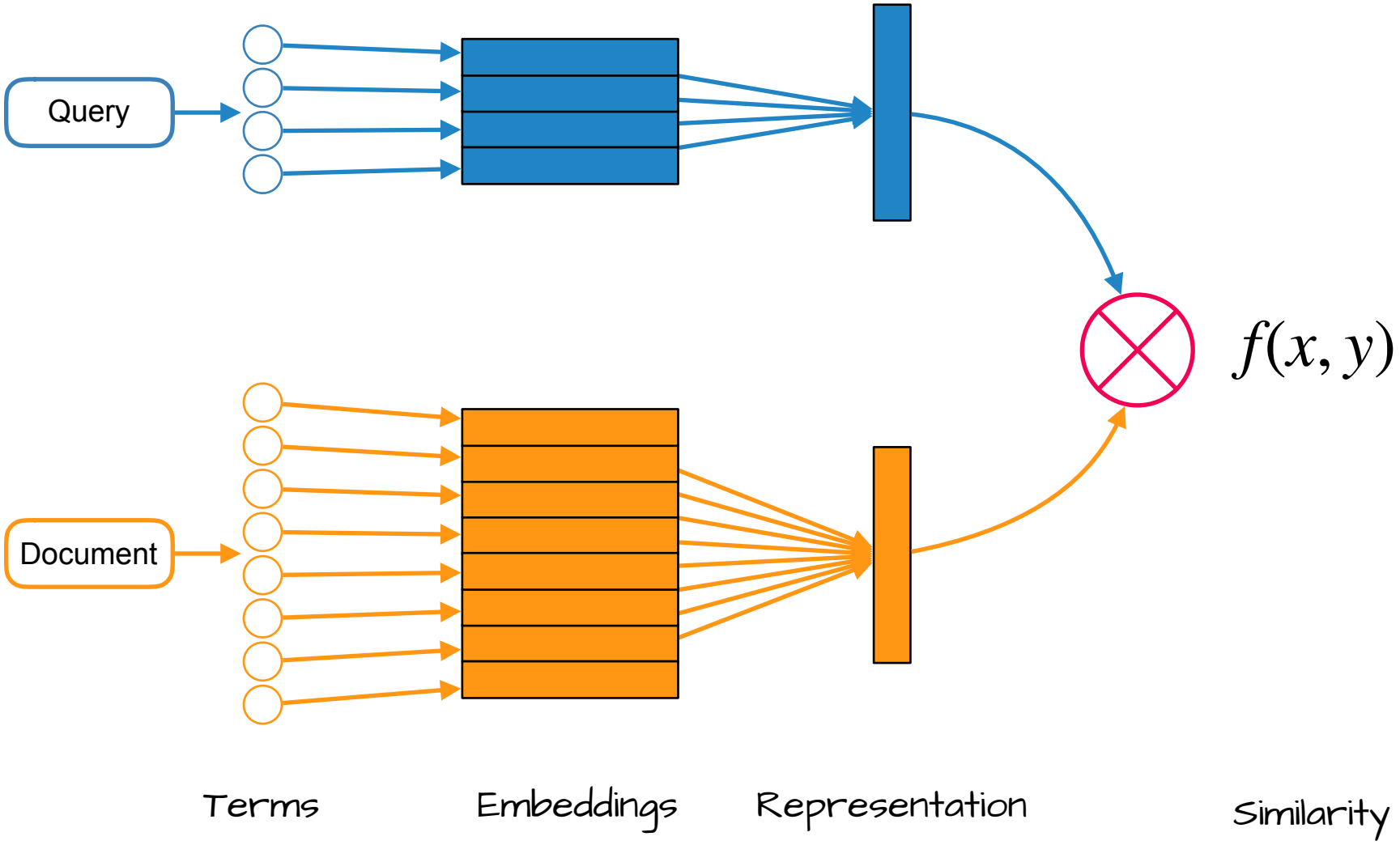
https://arxiv.org/pdf/1707.07700.pdf

Representation focused models



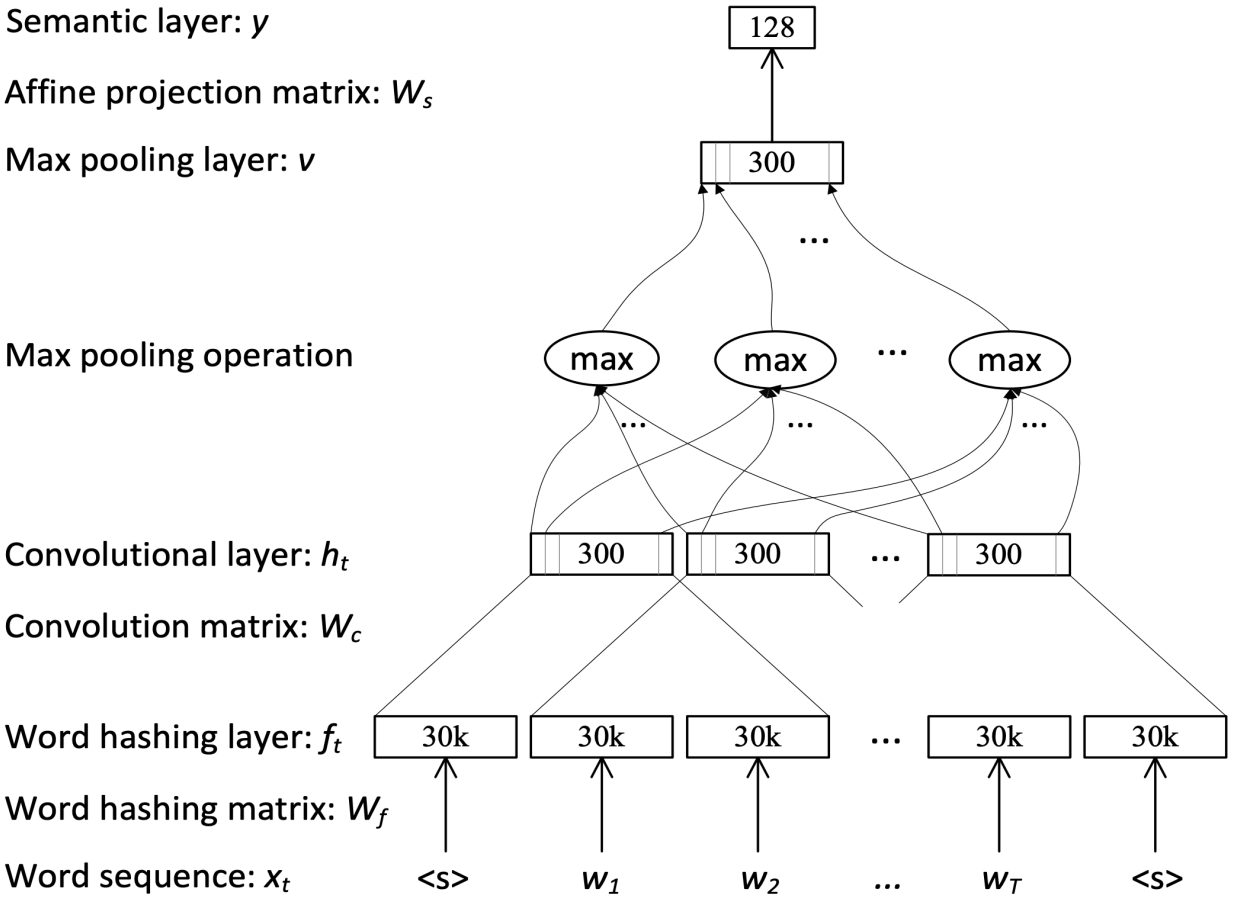
https://arxiv.org/pdf/1707.07700.pdf

Representation focused models



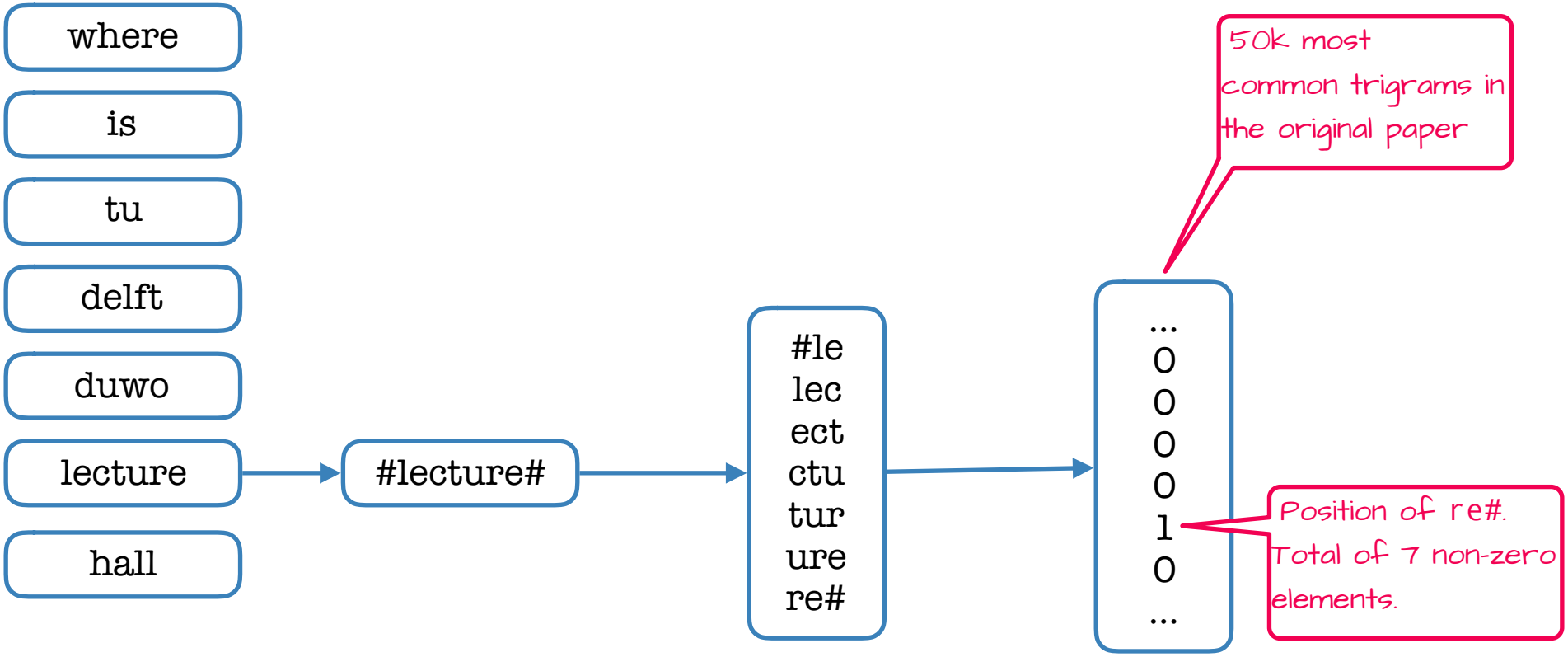
https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/www2014_cdssm_p07.pdf

CDSSM - Convolutional Deep Structure Semantic Models



https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/www2014_cdssm_p07.pdf

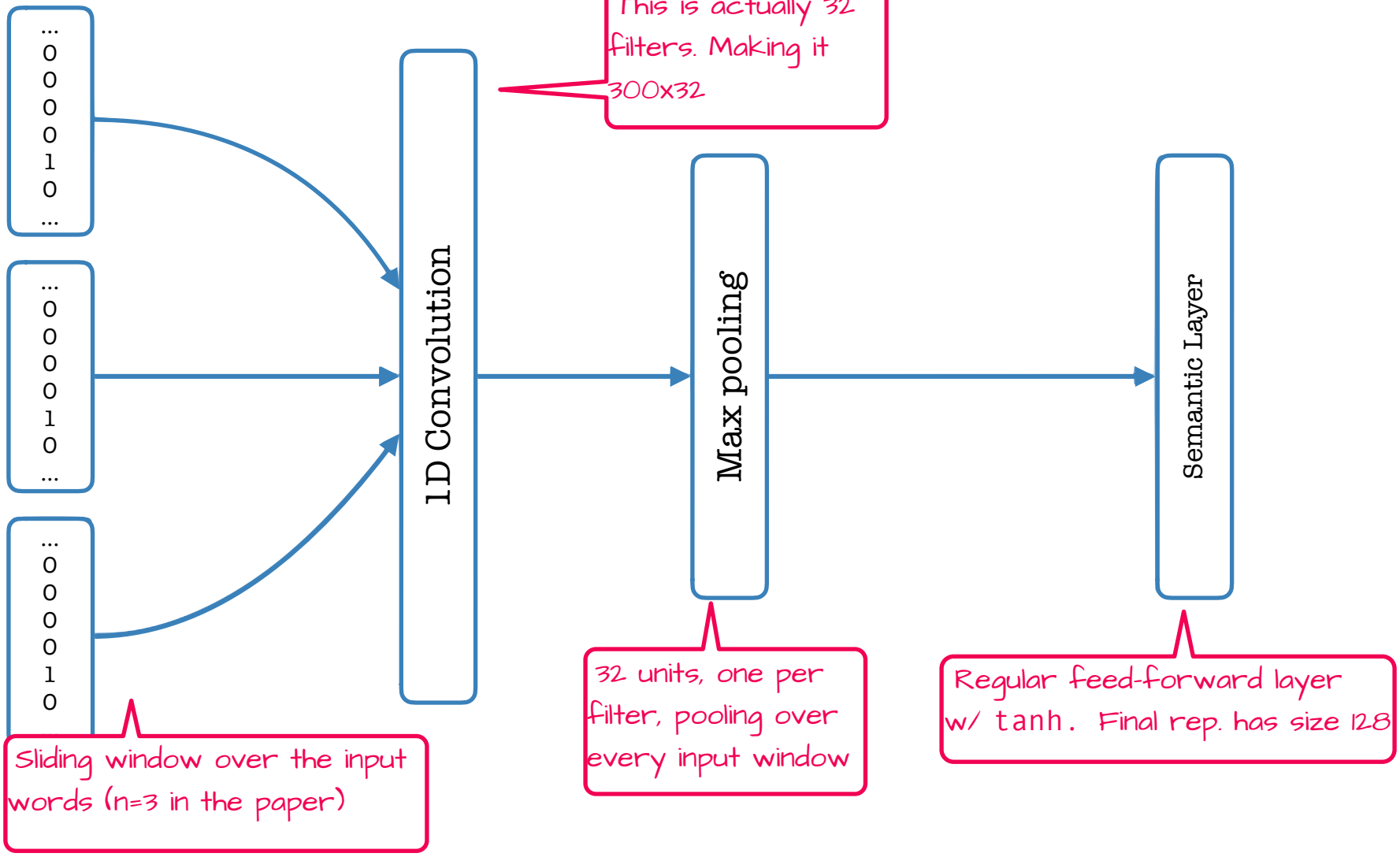
CDSSM - Word Hashing Layer



Input query | End-of-word markers | Trigrams | trigram feature vector

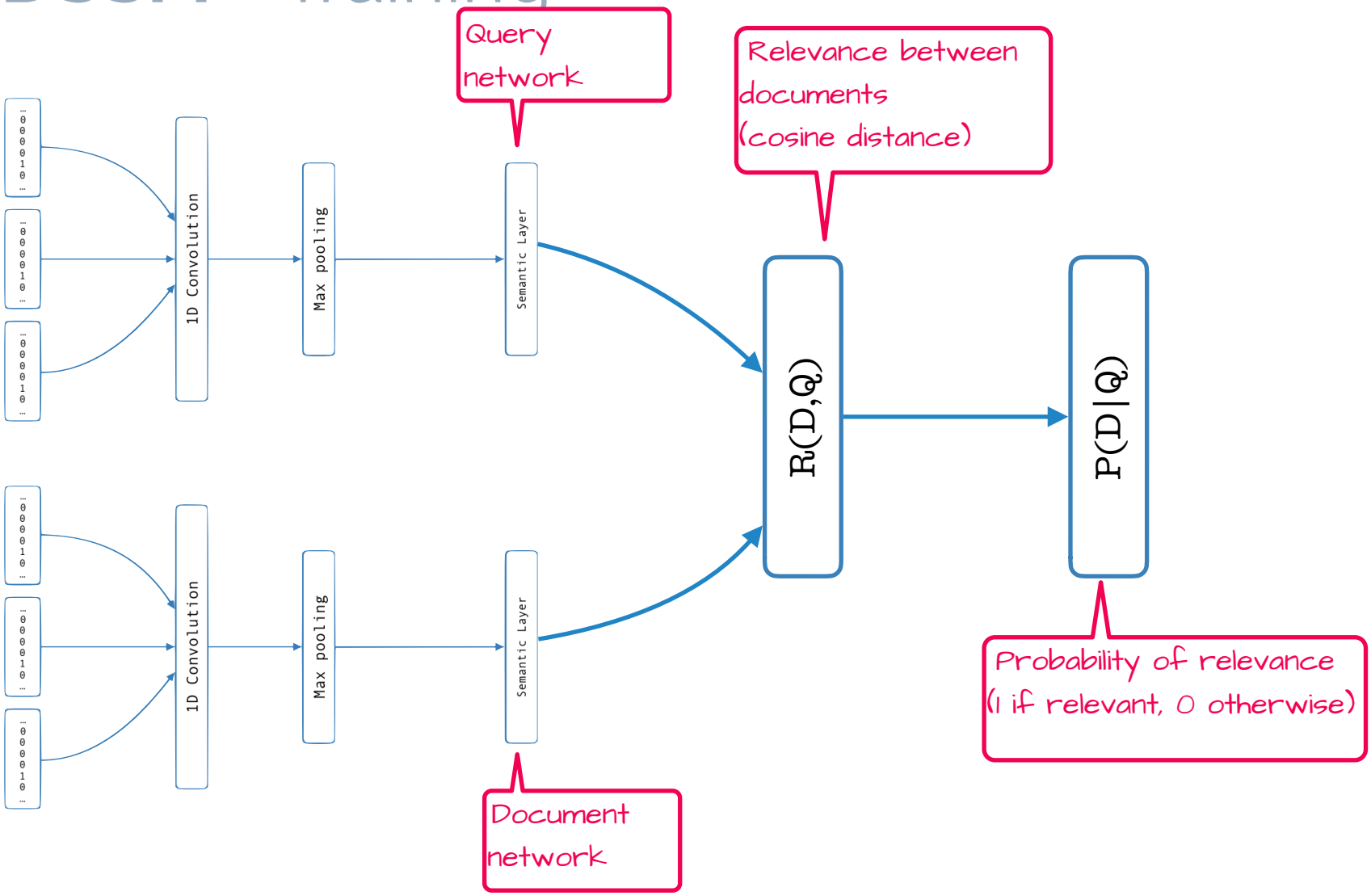
https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/www2014_cdssm_p07.pdf

CDSSM - Convolutional Layer



https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/www2014_cdssm_p07.pdf

CDSSM - Training



https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/www2014_cdssm_p07.pdf

CDSSM - Results

Using bing queries

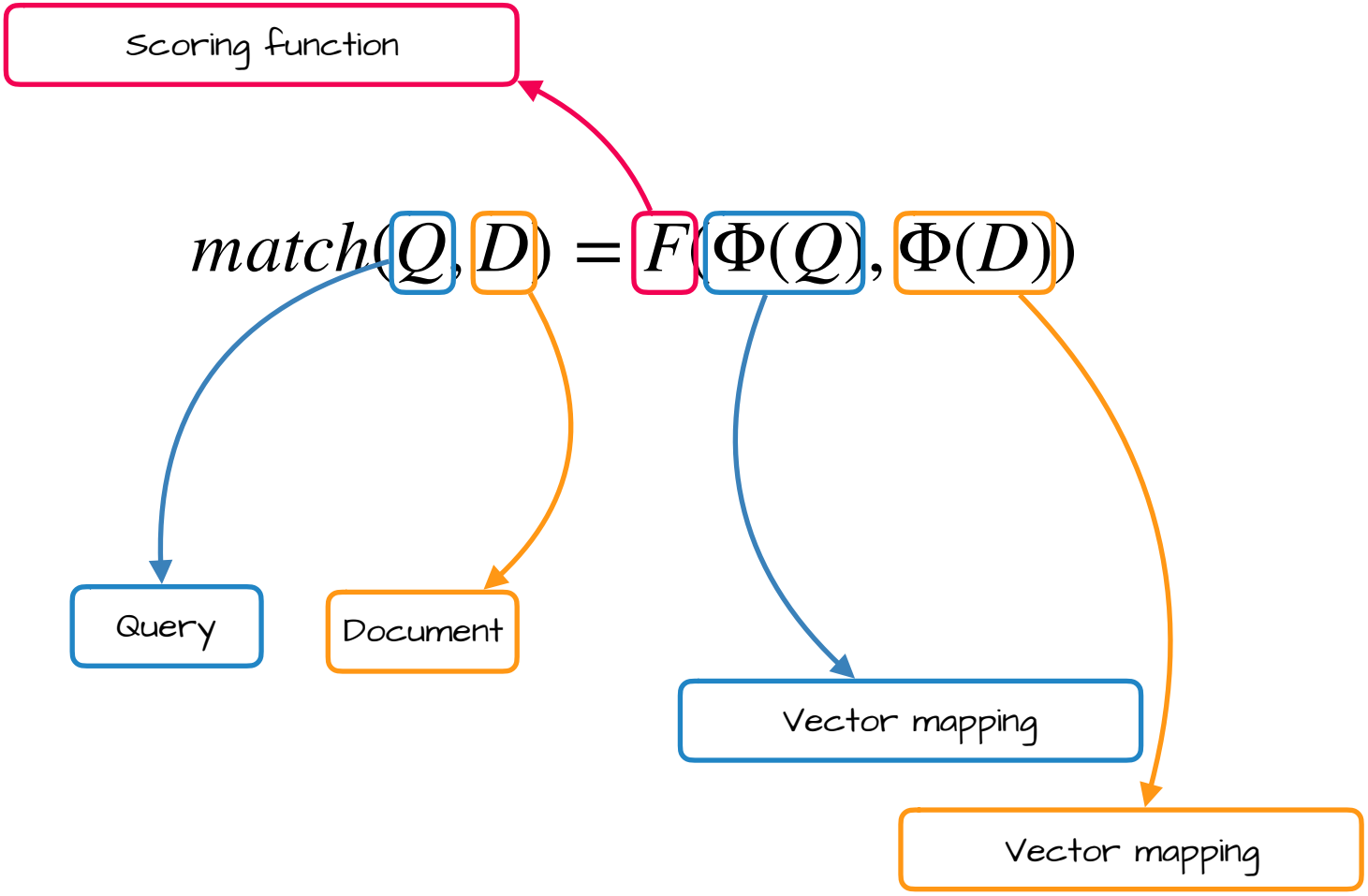
#	Models	NDCG@1	NDCG@3	NDCG@5
1	BM25	0.305	0.328	0.388
2	ULM	0.304	0.327	0.385
3	WTM	0.315	0.342	0.411
4	PTM (len <=3)	0.319	0.347	0.413
5	DSSM	0.320	0.355	0.431
6	**C-DSSM (win=3)**	**0.342**	**0.374**	**0.447**

https://dl.acm.org/citation.cfm?id=2767830

Interaction focused models

https://arxiv.org/pdf/1707.07700.pdf

Interaction-focused models



https://arxiv.org/pdf/1707.07700.pdf

Interaction-focused models

$$match(Q, D) = F(\Phi(Q), \Phi(D))$$

$$H \otimes G \otimes M(\Phi(Q)\Phi(D))$$

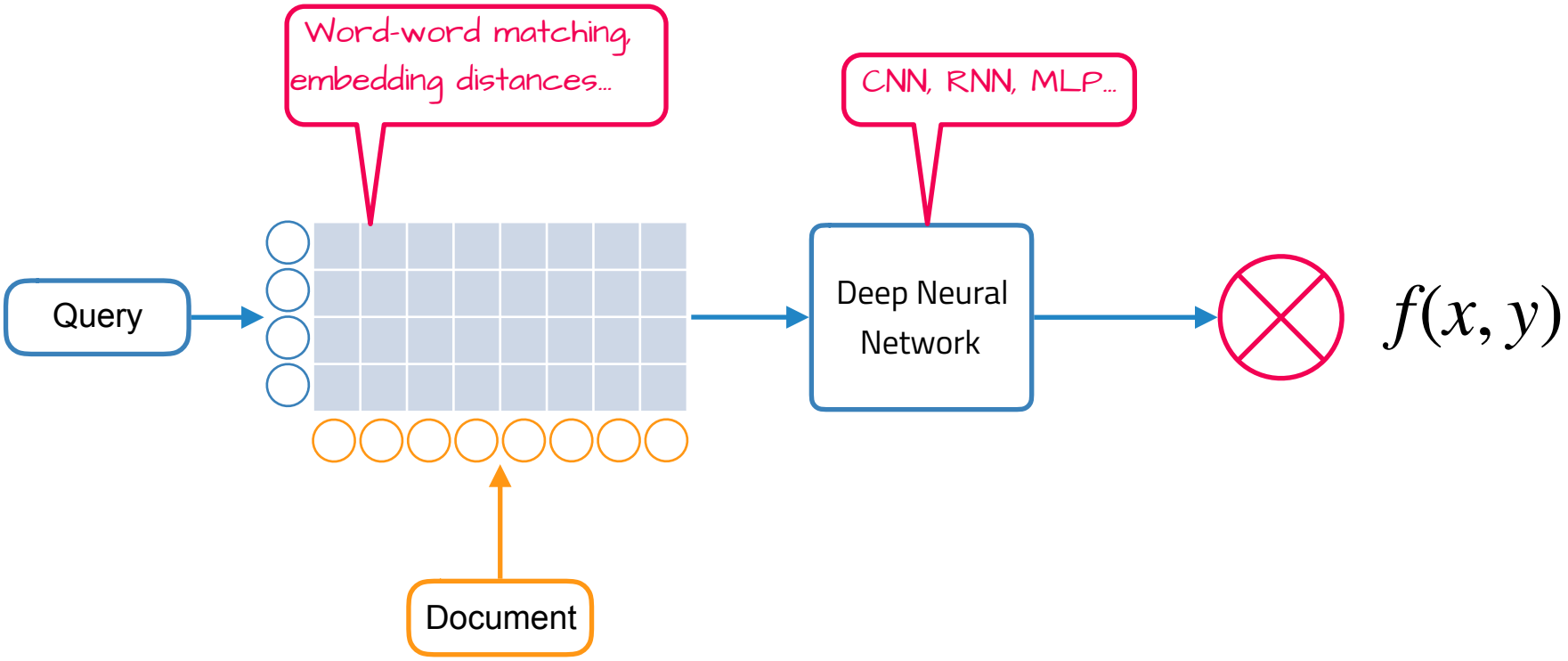
Local interactions between query and document

Relevance function

Deep neural network

https://arxiv.org/pdf/1707.07700.pdf

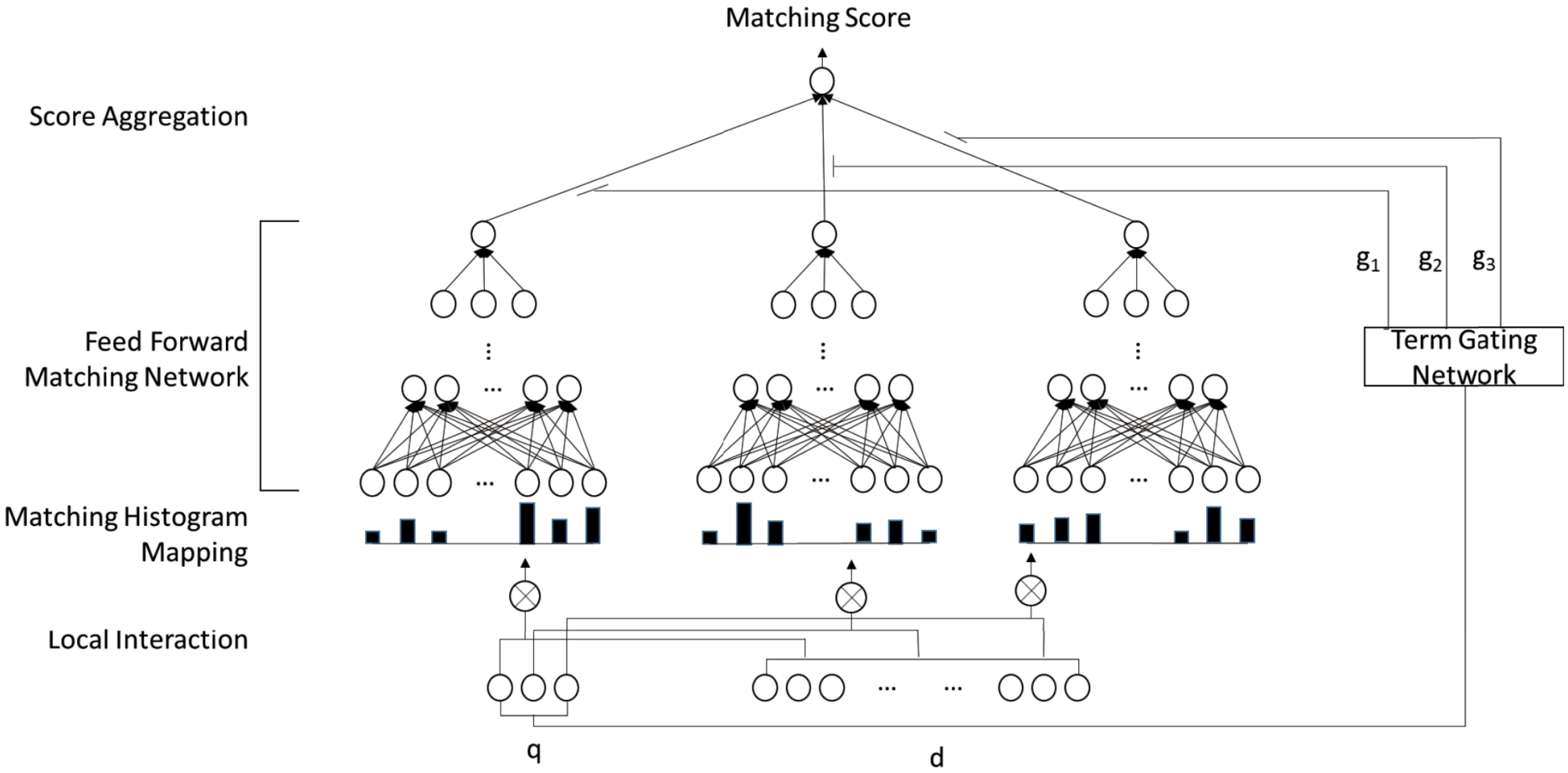
Interaction-focused models



Interaction matrix (M) Deep Neural Net (G) Relevance Score (H)

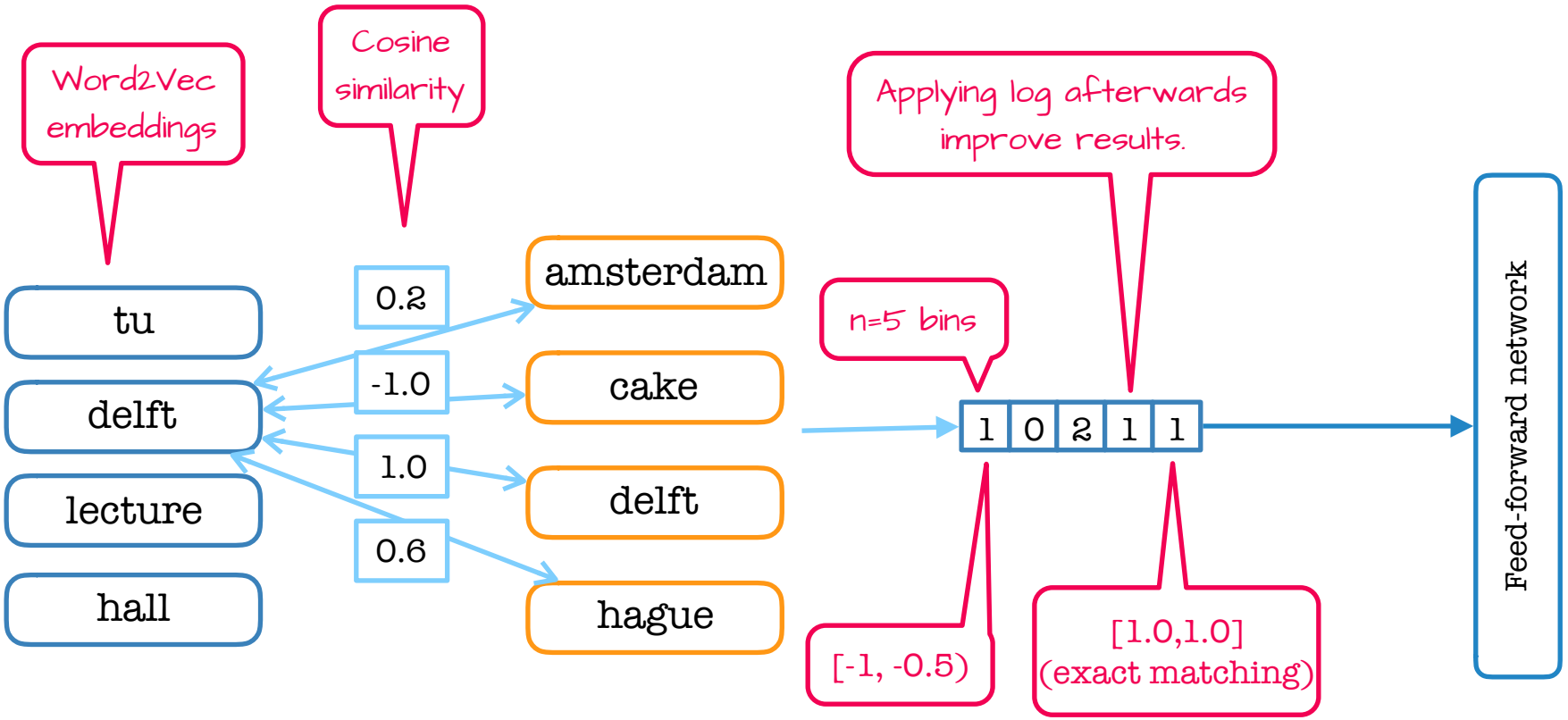
https://arxiv.org/pdf/1711.08611.pdf

DRMM - Deep Relevance Matching Model



https://arxiv.org/pdf/1711.08611.pdf

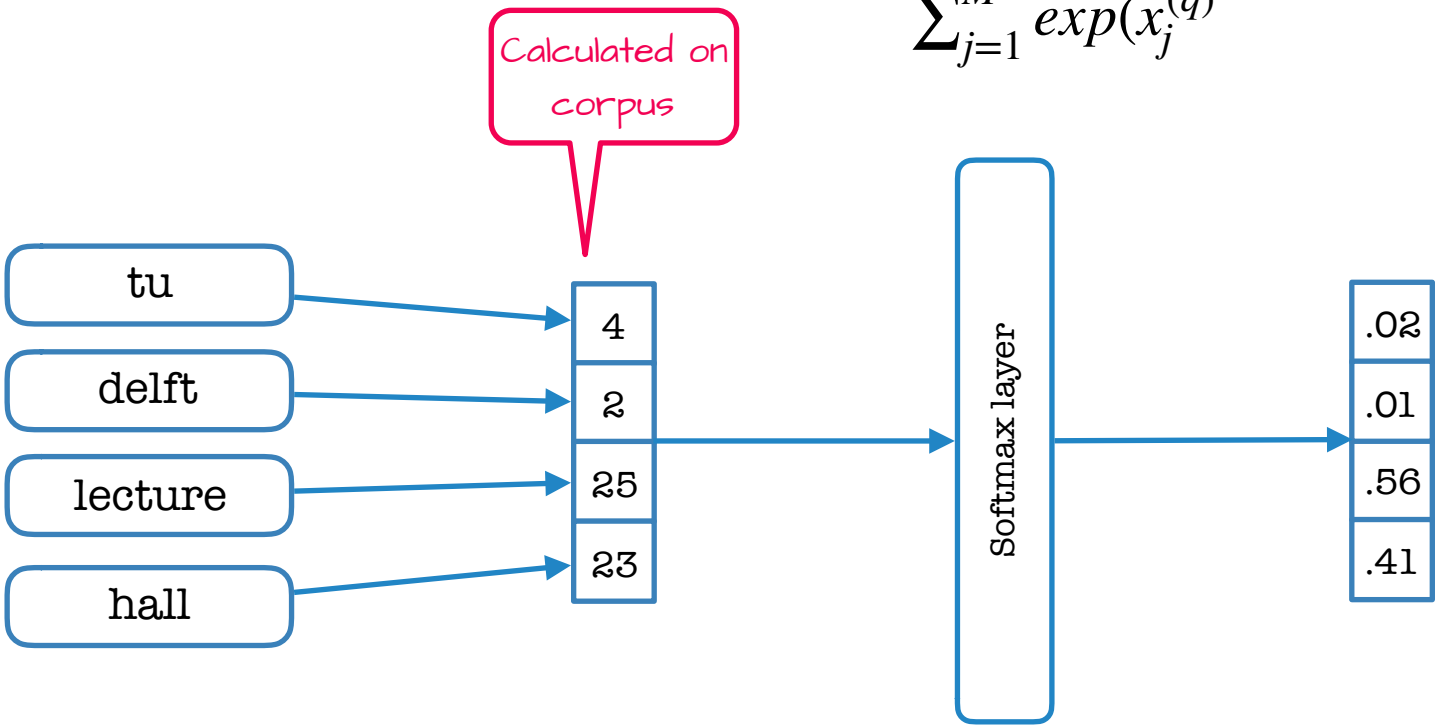
DRMM - Local Interactions and Histogram Mapping



https://arxiv.org/pdf/1711.08611.pdf

DRMM - Term Gating Network

$$g_i = \frac{exp(x_i^{(q)})}{\sum_{j=1}^{M} exp(x_j^{(q)})}$$

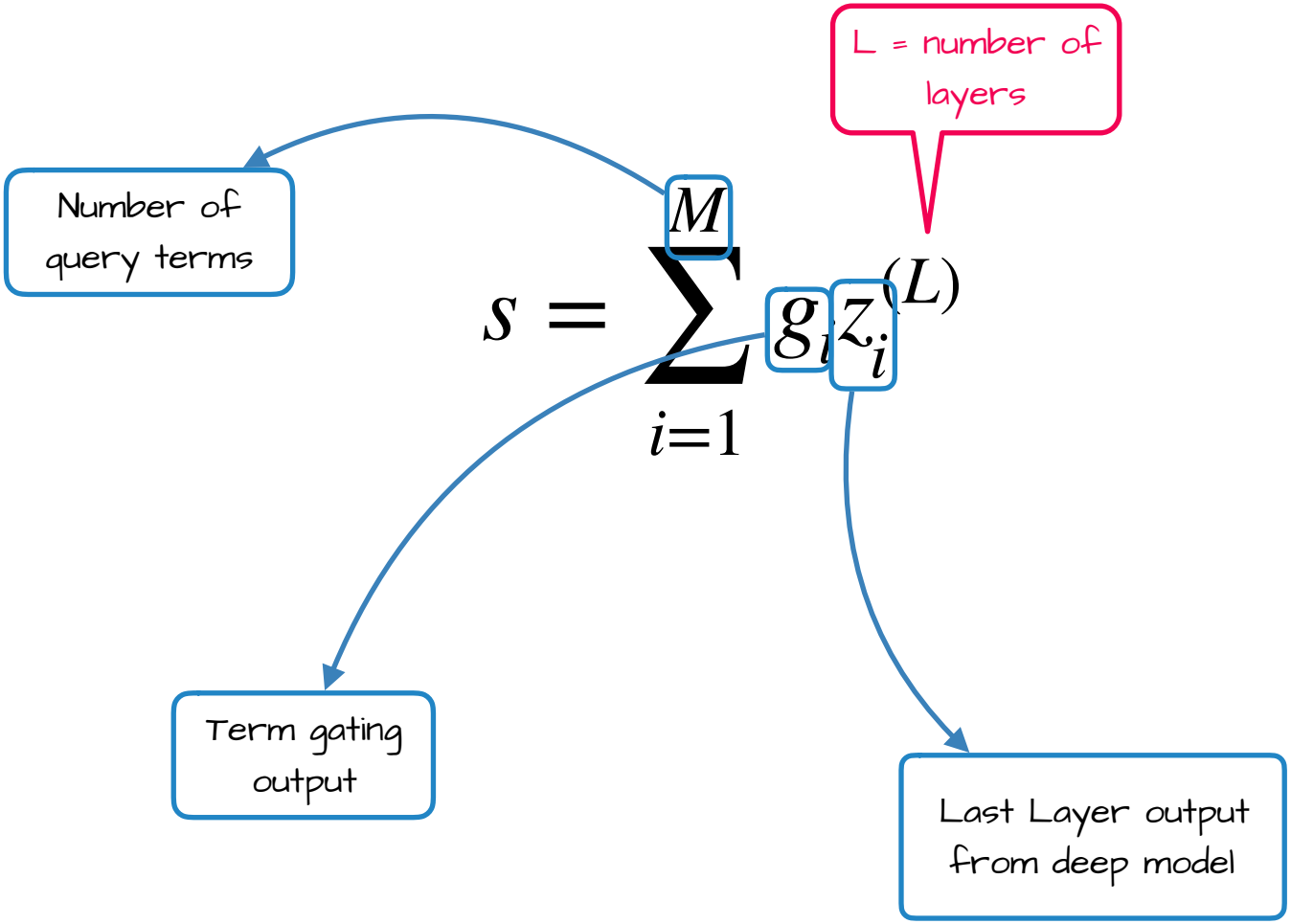


Input query

IDF Vector

https://arxiv.org/pdf/1711.08611.pdf

DRMM - Final aggregation



https://arxiv.org/pdf/1711.08611.pdf

DRMM - Results

Robust04, titles only

#	Models	MAP	NDCG@20	P@20
1	QL	0.253	0.415	0.369
2	BM25	0.255	0.418	0.370
3	CDSSM	0.067	0.146	0.125
4	ARC-II	0.067	0.147	0.128
5	**DRRM**	**0.279**	**0.431**	**0.382**

https://dl.acm.org/citation.cfm?id=3052579

Duet: local and distributed representations

https://dl.acm.org/citation.cfm?id=3052579

We need both exact term matches and semantic matches

Local representation

The President of the United States of America (POTUS) is the elected head of state and head of government of the United States. The president leads the executive branch of the federal government and is the commander in chief of the United States Armed Forces. Barack Hussein Obama II (born August 4, 1961) is an American politician who is the 44th and current President of the United States. He is the first African American to hold the office and the first president born outside the continental United States.

Q="united states president"

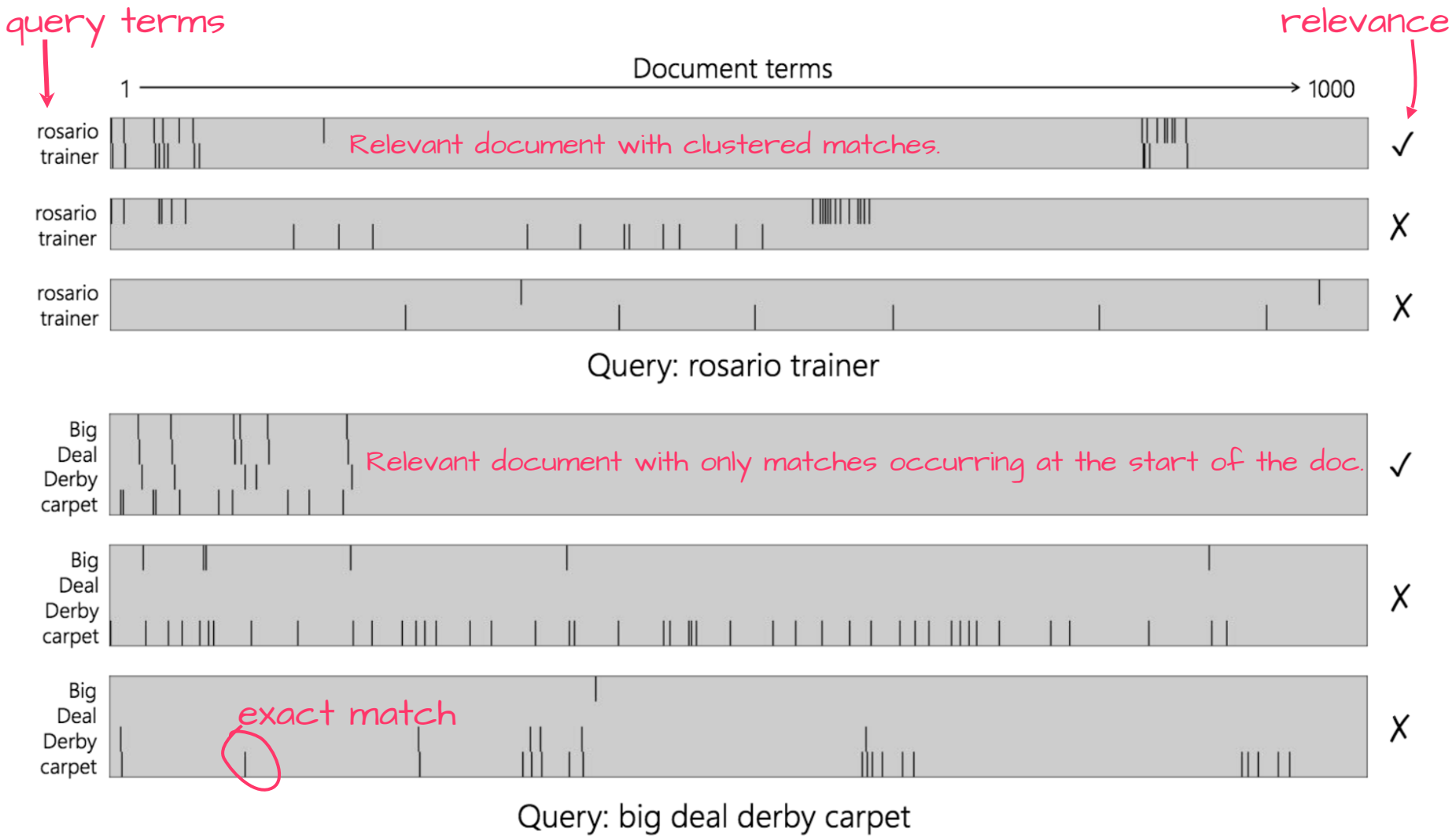
Shade of green indicates the drop in retrieval model's document score by individually removing each of the passage terms.

Distributed representation

The President of the United States of America (POTUS) is the elected head of state and head of government of the United States. The president leads the executive branch of the federal government and is the commander in chief of the United States Armed Forces. Barack Hussein Obama II (born August 4, 1961) is an American politician who is the 44th and current President of the United States. He is the first African American to hold the office and the first president born outside the continental United States.

https://dl.acm.org/citation.cfm?id=3052579

Patterns of query term matches



https://dl.acm.org/citation.cfm?id=3052579

Architecture

$$f(\mathbf{Q}, \mathbf{D}) = f_{local}(\mathbf{Q}, \mathbf{D}) + f_{distr}(\mathbf{Q}, \mathbf{D})$$

$$where\ \mathbf{Q} = [q_1, .., q_{n_q}]$$

$$and\ \mathbf{D} = [d_1, ..., d_{n_d}];$$

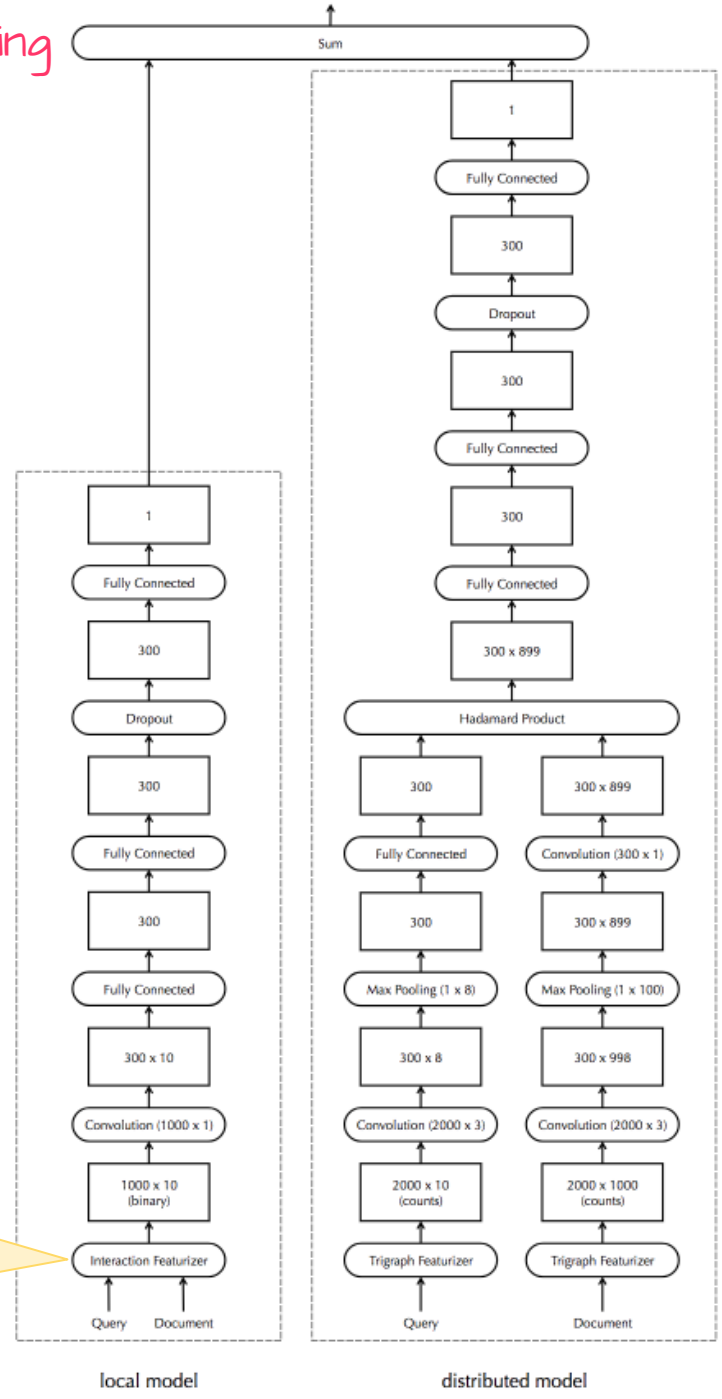
$$a\ term\ is\ represented\ by\ a\ m \times 1\ vector$$

Fixed input length: first 10 terms per query and first 1000 terms per document. Shorter queries/ docs are padded with zeros.

Local model takes one-hot encoding as input. Distributed model takes n-graph based representatio input

(2000 most frequent uni/bi/tri/4/5-grams).

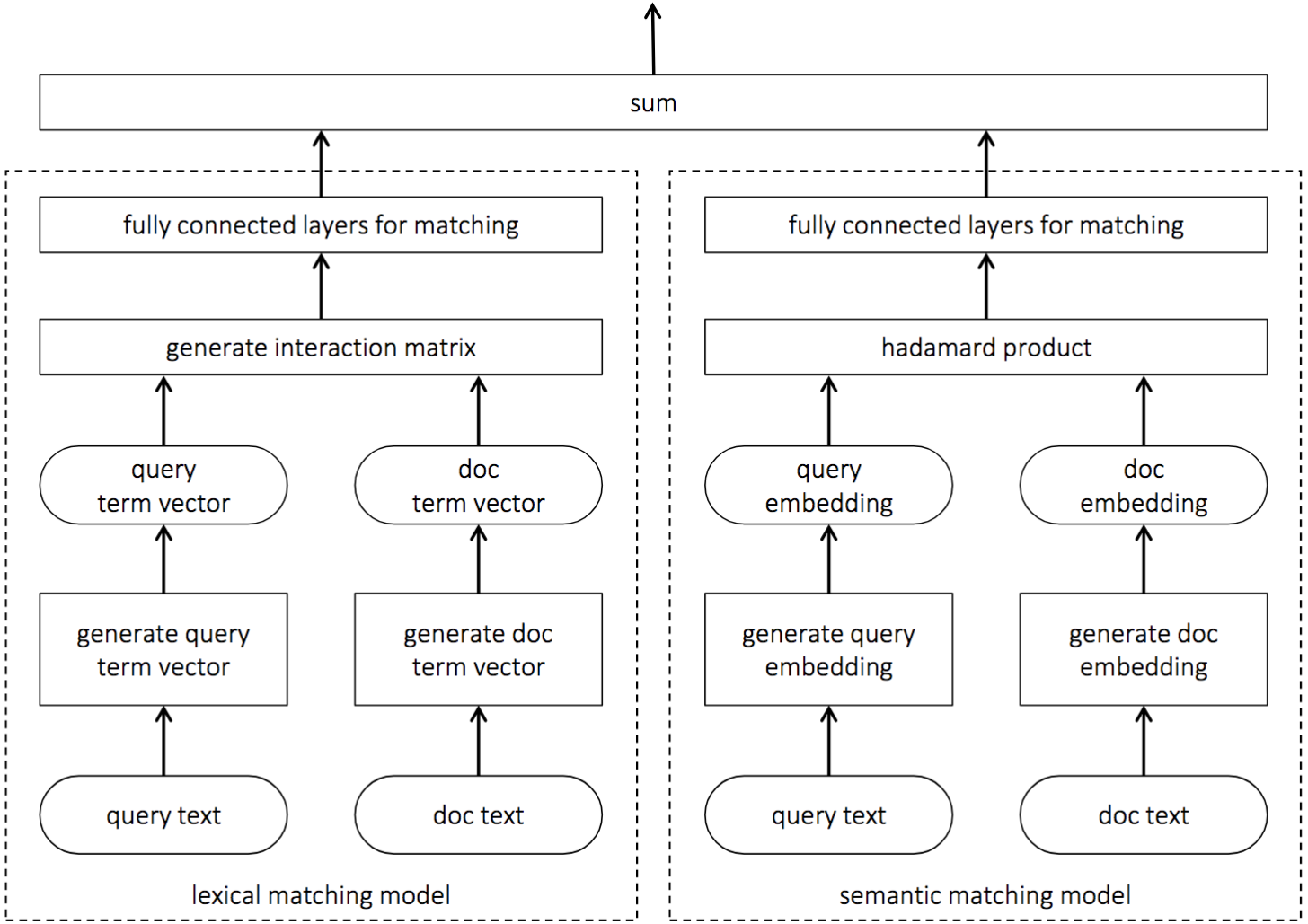
joint training



Patterns of term matches

https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

Architecture (the bigger picture)



https://dl.acm.org/citation.cfm?id=3052579

Experiments

Domain knowledge is key: "For training, we discarded all documents rated as perfect because a large portion of them fall under the **navigational intent,** which can be better satisfied by historical click based ranking signals"

Training data: **200K queries** sampled from Bing's search logs.

Approximately five documents were judged by human judges as either *perfect*, *excellent*, *good*, *fair*, or *bad*.

Test data: **8K queries**, with approx. 25 documents per query (**14%** of queries also occurred in the training set)

Four negative documents sampled per relevant document.

	NDCG@1	NDCG@10
Non-neural baselines		
LSA	22.4	44.2
BM25	24.2	45.5
DM	24.7	46.2
QL	24.6	46.3
Neural baselines		
DRMM	24.3	45.2
DSSM	25.8	48.2
CDSSM	*27.3*	48.2
DESM	25.4	*48.3*
Our models		
Local model	24.6	45.1
Distributed model	28.6	50.5
Duet model	**32.2**	**53.0**

https://dl.acm.org/citation.cfm?id=2767830

Caveats?

https://cs.uwaterloo.ca/~jimmylin/publications/Lin_SIGIRForum2018.pdf

A properly tuned BM25+RM3…

Condition	AP	P20
BM25	0.238	0.354
BM25 + Features	0.250	0.367
Neural_x	0.258	0.372
Neural_y	0.256	0.370
$\text{Neural}_x + \text{Neural}_y$		0.373
$\text{A} + \text{Neural}_y$		0.380
$\text{A} + \text{Neural}_y + \text{M}$		0.380
$\text{B} + \text{Neural}_y$	0.270	0.383
$\text{B} + \text{Neural}_y + \text{M}$	0.272	0.386
Anserini: QL	0.2481	0.3517
Anserini: BM25	0.2528	0.3598
Anserini: BM25 + RM3 (independent)	0.2991	0.3901
Anserini: BM25 + RM3 (joint)	0.2956	0.3931

This is a variation over DRMM…

Maybe that's why DL is not taking over IR yet…

http://rali.iro.umontreal.ca/rali/sites/default/files/publis/p59-nie%5B1%5D.pdf

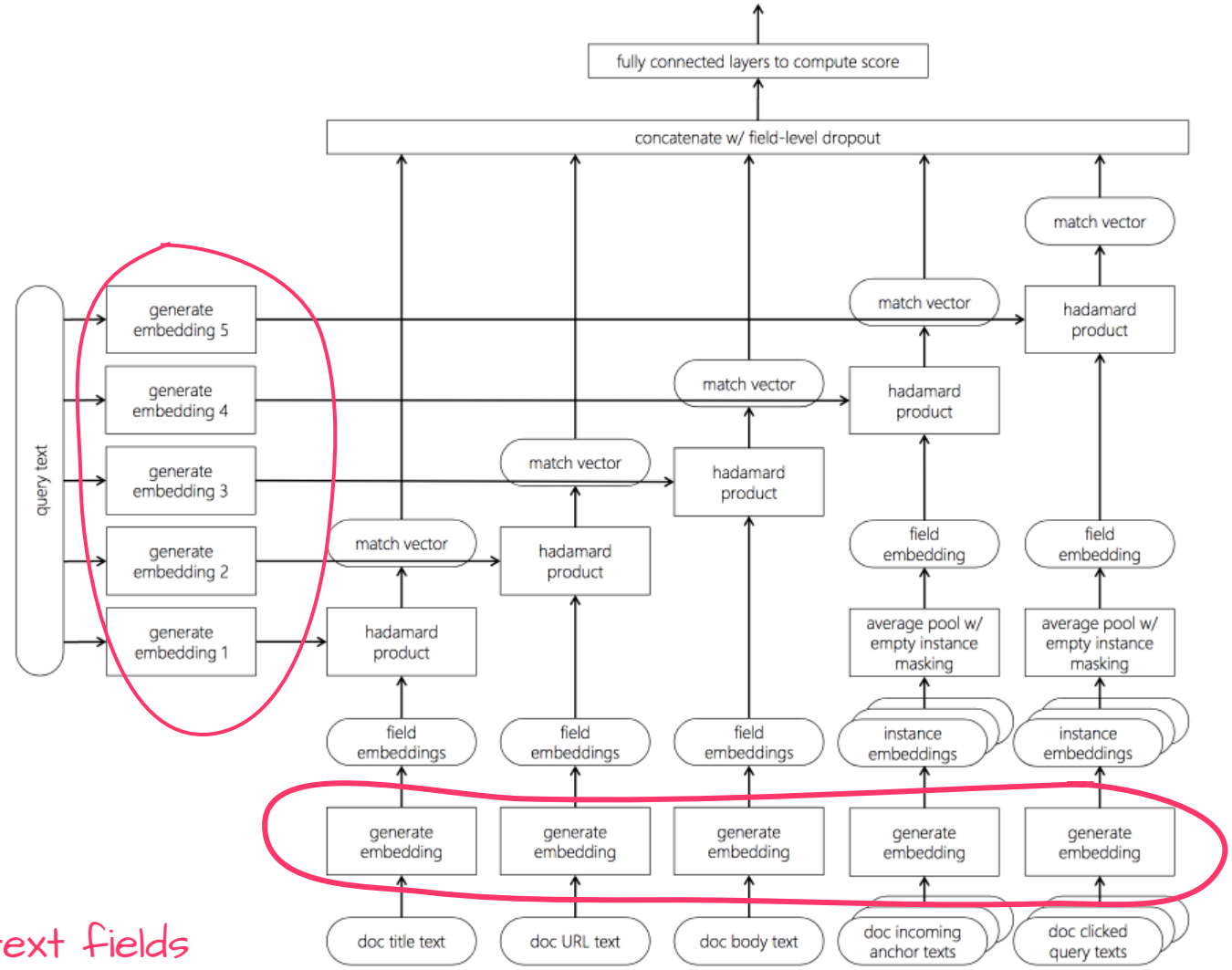
Insights

- Generally, interaction-focused techniques work best (Nie, et al., ICTR ’18).
- For short texts, vocabulary mismatch is more serious than for long texts
- MatchZoo has lots of models implemented in Python/TensorFlow (**BEWARE! Implementations may be different from original paper**!) [https://github.com/NTMC-Community/MatchZoo/].
- Properly turned “classical" IR models are still good (if not better than neural) at most IR tasks.

What now?

https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

Designing more architectures is not a problem ...



text fields

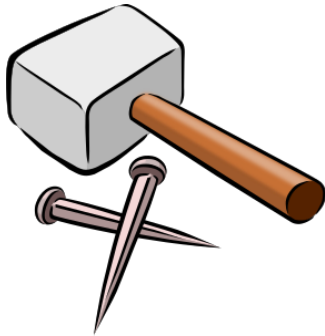
Learning embeddings separately performs better than one common embeddings space

https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/fntir-neuralir-mitra.pdf

However, given the pace at which the area of deep learning is growing [...] we should be wary of the combinatorial explosion of trying every model on every IR task.

We should not disproportionately focus on maximizing quantitative improvements and in the process neglect theoretical understanding and qualitative insights. [...]

Neural models should not be the hammer that we try on every IR task, or we may risk reducing every IR task to a nail.



Many worthwhile issues to tackle

- Lack of large-scale training data for supervised DNN remains the biggest obstacle (TREC-style evaluations will never scale up, we cannot leave the field to industry); **weak supervision** has attracted a lot of attention recently

- So far, most DNNs have been applied to ad hoc retrieval and Web search, but of course IR is much bigger than that

- Clever tricks from the machine learning community are not yet often used (e.g. **curriculum learning**)

- **Reproducibility** and benchmarking

- **Metrics** (of course!) for new tasks, e.g. answer generation

- Instead of applying DNN building blocks, we should **develop our own** (relevance, IDF, …)

That's it for today!

Slack: in4325_2019.slack.com

Email: in4325-ewi@tudelft.nl

MSc project ideas? A.BarbosaCamara@tudelft.nl