Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding
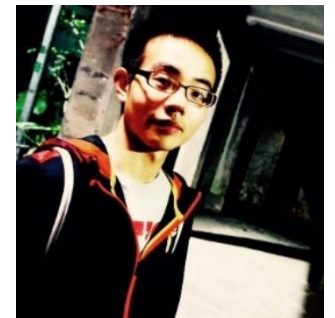
阅读报告

张继元
201728014628057
自动化研究所
zhangjiyuan2017@ia.ac.cn

Chenyan Xiong, Russell Power and Jamie Callan. *Explicit Semantic Ranking for Academic Search via Knowledge Graph Embedding*. WWW 2017.

Challenge

Exact matching query and document's words does not solve IR.

Bag-of-words representation

↓
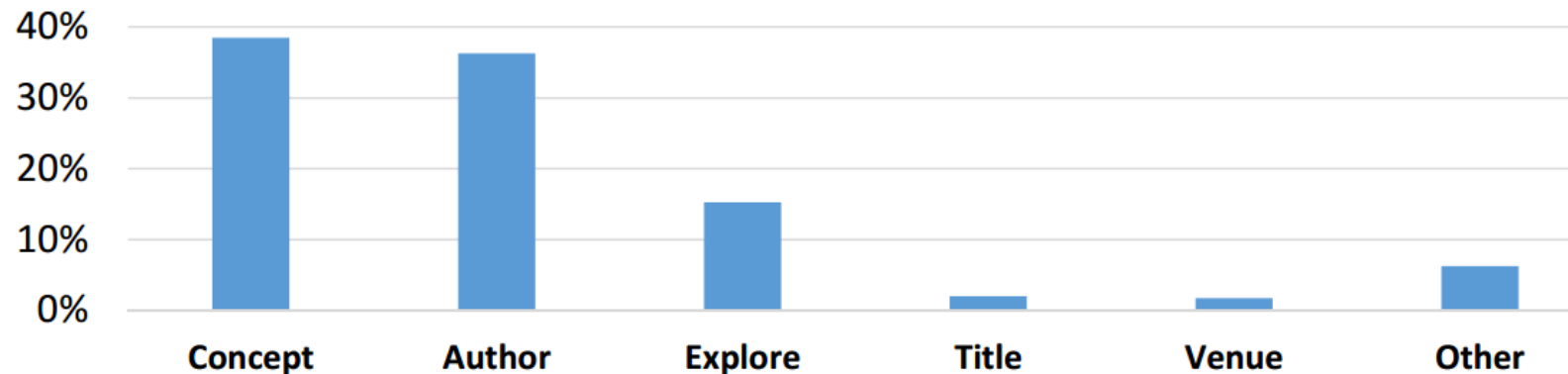
Vocabulary Mismatch

Query

> CMU location

Document

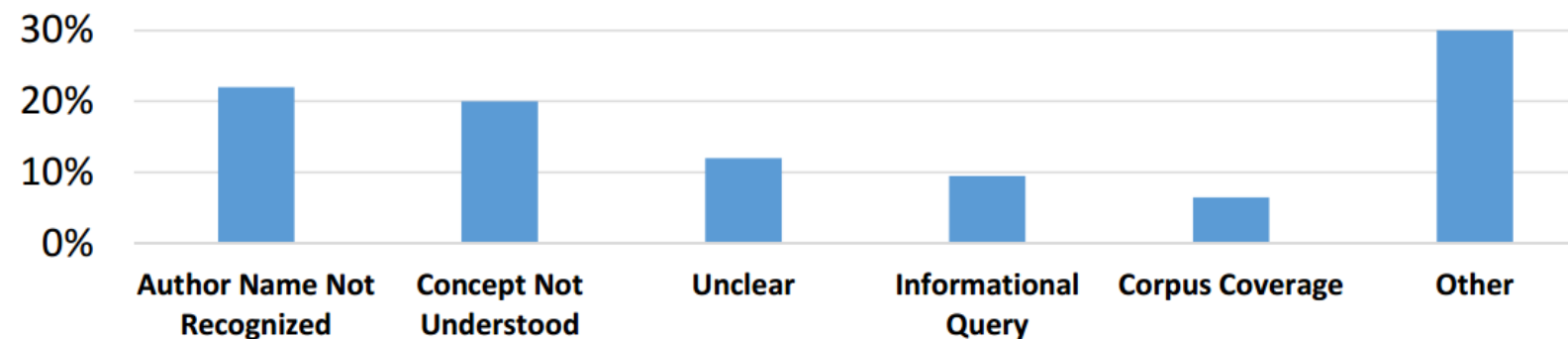> Carnegie Mellon University is a private research university in …

Query Log Analysis

(a) Query Intent Distribution

(b) Error Source Distribution

Contribution

1. An in-domain knowledge graph construction for academic search

2. Bag-of-Entities based text representation

3. Explicit Semantic Ranking

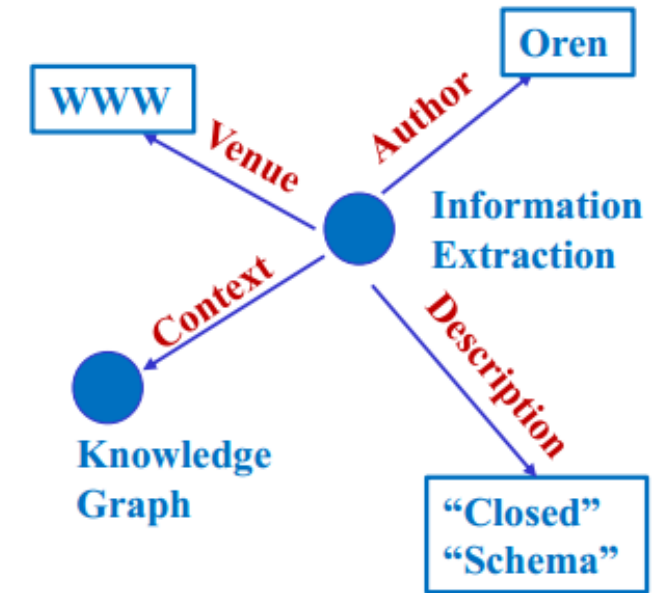# In-domain knowledge graph construction

Entities are from a general domain knowledge graph (Freebase):

- Very good coverage in academic search

- Better than automatic key phrase extraction

Bag-of-Entities based text representation

- **Domain specific edges**

1. Author Edge
2. Venue Edge
3. Context Edge
4. Description Edge

- **An entity embedding model for each edge type**

  Skip-gram model on each Entity →

  Author | Venue | Context | Description Pair
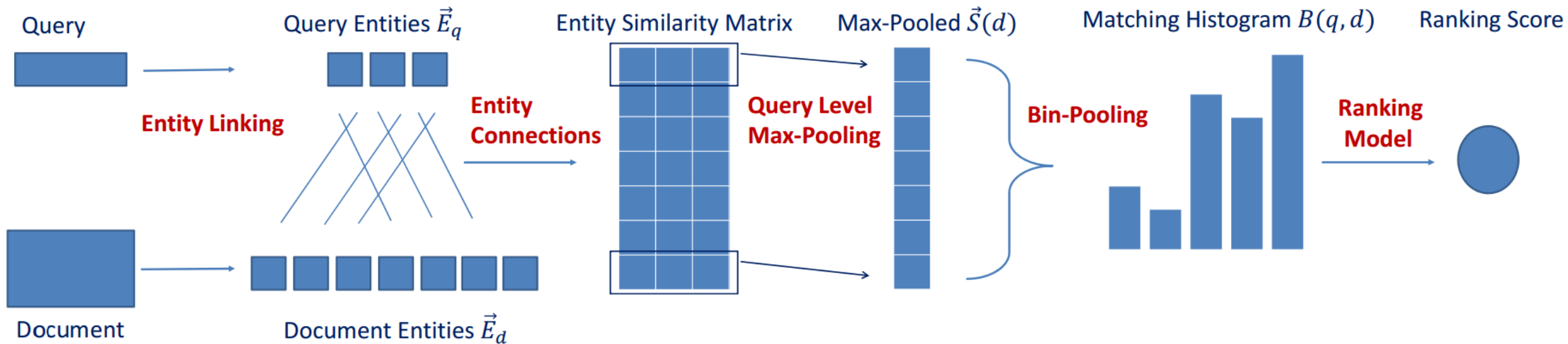
Explicit Semantic Ranking

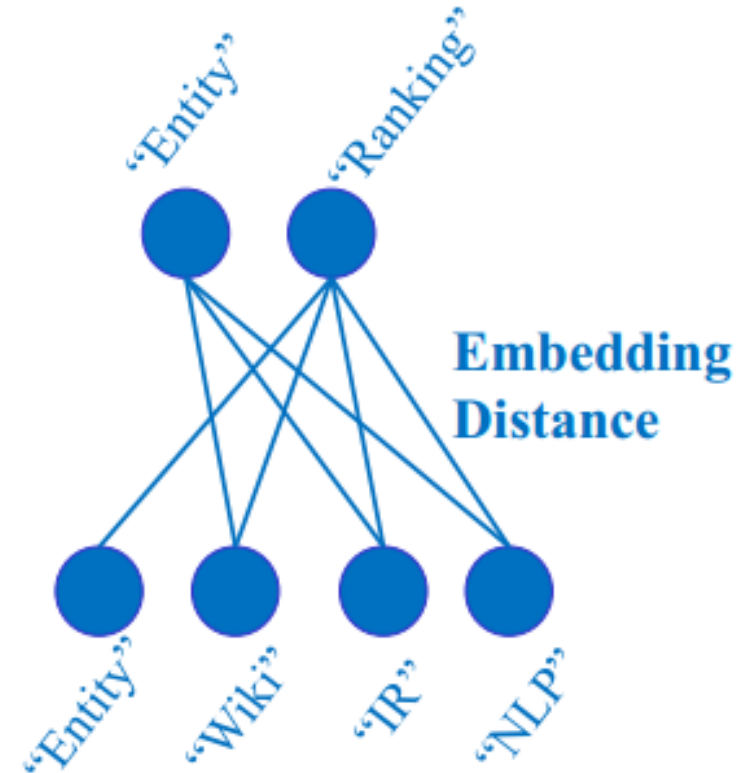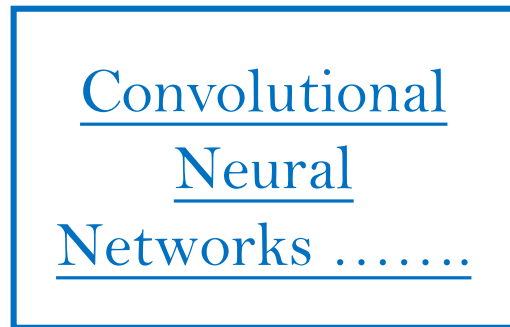Figure 2: The Framework of Explicit Semantic Ranking (ESR).

Advantages

1. **Smart Phasing**
2. **Embedding-based Soft Match**

Query

CNN

Document

Convolutional
Neural
Networks ......

"Entity" "Ranking"

**Embedding
Distance**

"Entity" "Wiki" "IR" "NLP"

Evalution Results

Method	NDCG@1		NDCG@5		NDCG@10		NDCG@20		W/T/L
tf.idf-F	0.2020	-59.65%	0.2254	-54.93%	0.2741	-47.57%	0.3299	-39.91%	28/01/71
BM25-F	0.2512	-49.81%	0.2890	-42.20%	0.3150	-39.74%	0.3693	-32.75%	32/01/67
Semantic Scholar	0.5006	$-$	0.5000	$-$	0.5228	$-$	0.5491	$-$	$-/-/-$
ESR-Author	$0.5499^{\dagger}$	$+9.85\%$	$0.5501^{\dagger}$	$+10.02\%$	$0.5671^{\dagger}$	$+8.47\%$	$0.5935^{\dagger}$	$+8.08\%$	60/10/30
ESR-Context	$0.5519^{\dagger}$	$+10.25\%$	$0.5417^{\dagger}$	$+8.35\%$	$0.5636^{\dagger}$	$+7.80\%$	$0.5918^{\dagger}$	$+7.77\%$	58/04/38
ESR-Desc	0.5304	$+5.96\%$	$0.5496^{\dagger}$	$+9.92\%$	$0.5536^{\dagger}$	$+5.88\%$	$0.5875^{\dagger}$	$+6.99\%$	55/11/34
ESR-Venue	$\mathbf{0.5638}^{\dagger}$	$+12.63\%$	$\mathbf{0.5700}^{\dagger}$	$+13.99\%$	$\mathbf{0.5795}^{\dagger}$	$+10.83\%$	$\mathbf{0.6090}^{\dagger}$	$+10.91\%$	59/11/30
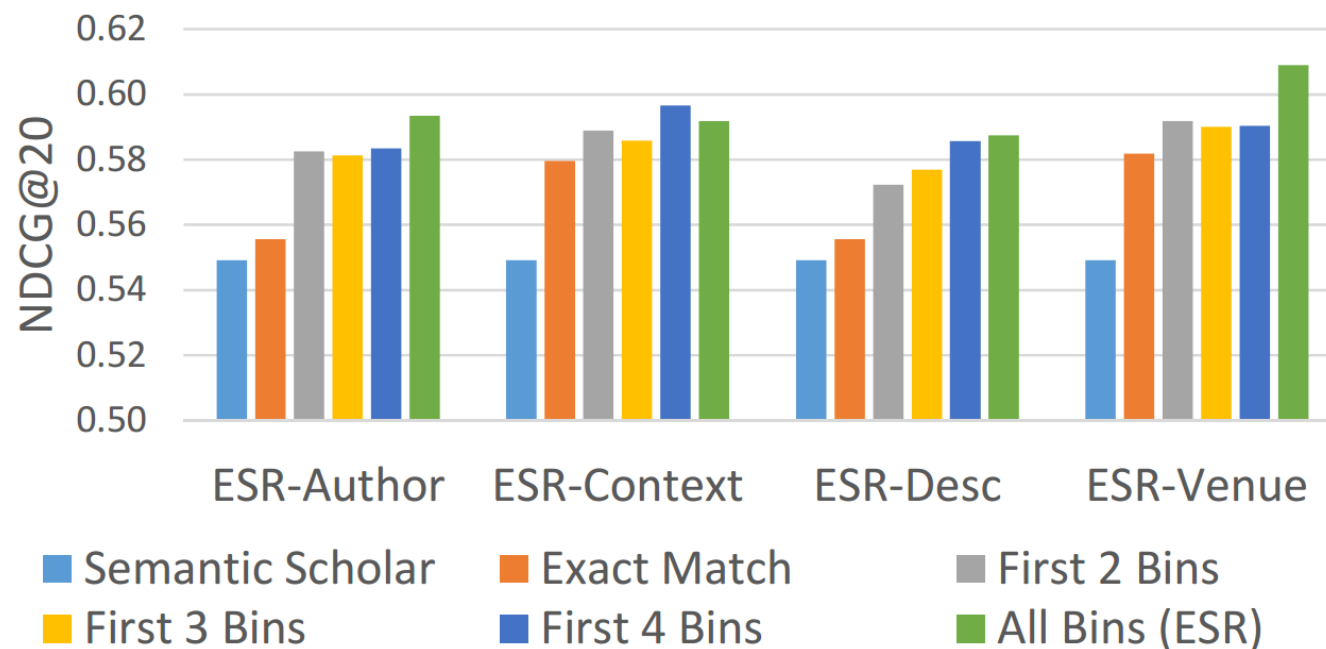
1、用最短也最通俗的一段话描述本论文内容(动机，方法，结果)

- 动机：改进词袋模型的局限性
- 方法：引入知识图谱的实体表示改进查询的匹配度
- 结果：明显提高查询的匹配度，特别是在基于词袋排序失败的一些查询中

2、本文的最大创新是什么？

①直接继承wiki或者freebase的图谱信息构建知识图谱，迅速建立当前领域的图谱。

②将知识图谱方法引入了检索领域，用bag-of-entity的表示形式，改变用词袋模型对查询和文档建模的方式，将语义信息引入到搜索中，既能实现exact-match，也能实现soft-match。

3、觉得本文有什么提高的空间？

构建bag-of-entity表示时候的edge type可以有更多的考虑，最后基于直方图的排序方法也可以有一些其他方法的尝试对比。

4、对自己的启发是什么？

对经典方法缺陷的反思，并能够结合最新的一些方法做改进。