Neural Query Expansion for Code Search

The 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages

Authors

Swarat Chaudhuri

Satish Chandra

CONTENTS

Background

NCS是一项代码搜索技术，采用自然语言查询并输出相关代码段

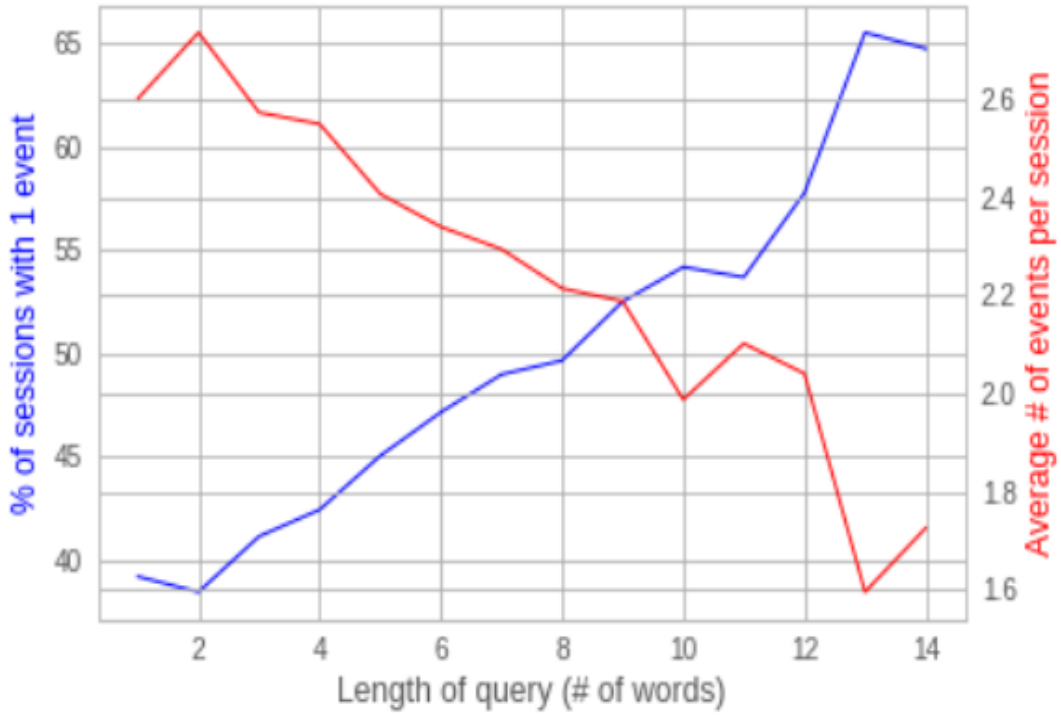
NCS存在的问题：

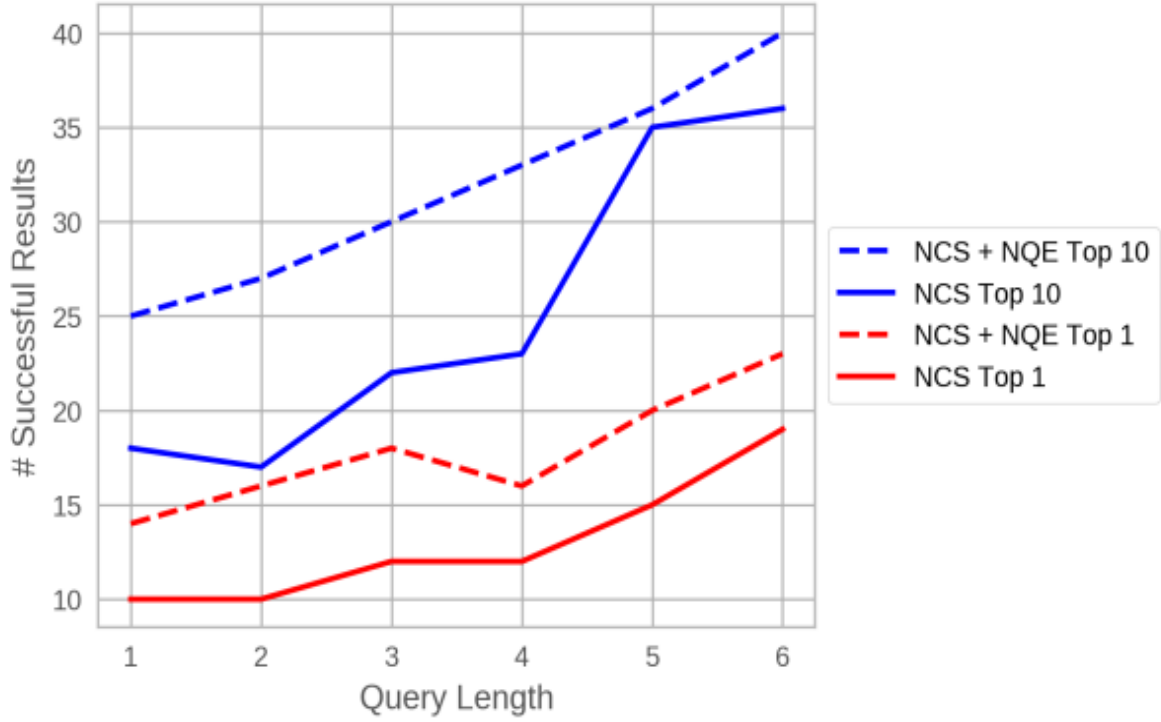
1）查询语句较短时NCS性能下降

2）查询语句较短时有更多的查询重新重写

3）查询语句较短时需要更多的时间浏览结果

Background


% of sessions with 1 event
Average # of events per session
Length of query (# of words)
65
60
55
50
45
40
26
24
22
20
18
16
2
4
6
8
10
12
14


Successful Results
Query Length
NCS + NQE Top 10
NCS Top 10
NCS + NQE Top 1
NCS Top 1
40
35
30
25
20
15
10
1
2
3
4
5
6

CONTENTS

Definition

D	程序的语料库
d	某一个具体的程序（$\mathrm{d} \in D$）
V_m	D中所有的方法名
split	一个基于Snake和Camel的切割函数
V_k	由Vm中的方法名经过split函数得到的关键词集合

$$V_k = \bigcup_{v \in V_m} split(v)$$

Definition

查询序列 $X = \{x_1, \ldots, x_n\}\ x_i \in V_k$

结果序列 $R = \langle d_1, \ldots, d_k \rangle\ d_X \in \mathcal{D}$

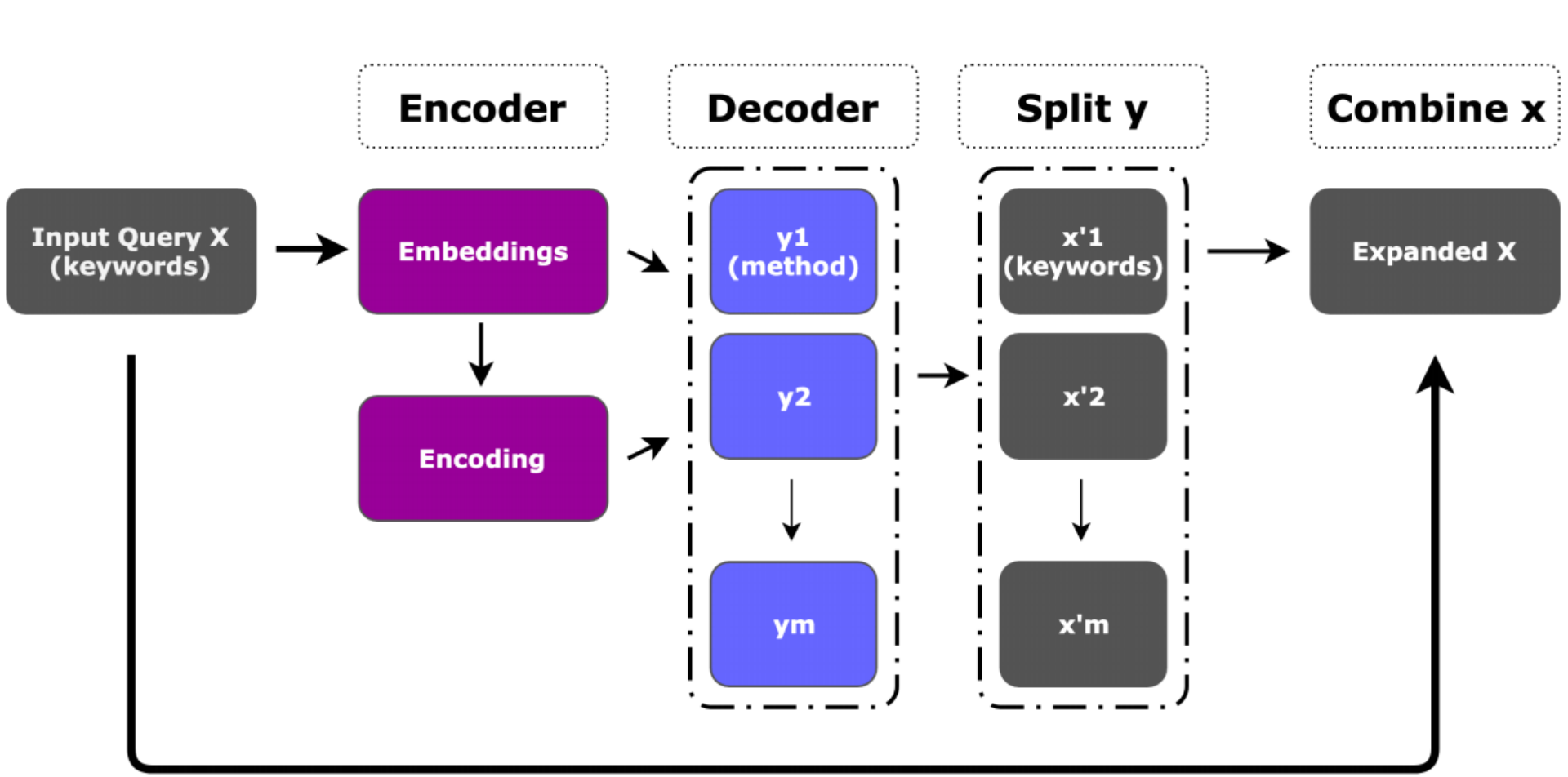
$rank(d_X, R) = \mathrm{i}$

$M : \varphi(v_k) -> S(D)$

$Q : \mathcal{P}(V_k) \rightarrow \mathcal{P}(V_k)$

$Y = \langle y_1, \ldots, y_m \rangle$

NQE


Encoder
Decoder
Split y
Combine x
Input Query X (keywords)
Embeddings
Encoding
y1 (method)
y2
ym
x'1 (keywords)
x'2
x'm
Expanded X

CONTENTS

Evaluation

Two instantiations： NCS and BM25

Another alternate query expansion model: FIM

Data Collection

737 public Android repositories(105,747 files 308,309 valid method bodies)

95%, 3%, and 2% for training, testing, and validation

GeneratingXquery from Y (TF-IDF dataset)

1. from d, the method calls are extracted to form Y.
2. take the top 50% TF-IDF methods from Y
3. extract the keywords
4. form candidate tokens(X_{cand})

Manually creating Xquery (Manual Dataset)

given Y, create an Xquery from lengths 1 to 6.

Results

RQ1: Does NQE improve performance for shorter queries?

Table 2. The number of Stack Overflow questions answered in the top 1, 5, 10 results with varying lengths of the queries. Search performance increases when *NCS* is aided by *NQE*, especially for shorter queries.

Top K	Query Length								
	1			4			All		
	NCS	NCS+ FIM	NCS+ NQE	NCS	NCS+ FIM	NCS+ NQE	NCS	NCS+ FIM	NCS+ NQE
1	10	10	**14**	12	12	**16**	20	18	**22**
5	15	15	**22**	20	21	**29**	33	28	**34**
10	18	18	**25**	23	25	**33**	40	40	40
	BM25	BM25+ FIM	BM25+ NQE	BM25	BM25+ FIM	BM25+ NQE	BM25	BM25+ FIM	BM25+ NQE
1	11	11	**17**	13	14	**16**	16	16	**18**
5	17	15	**20**	21	20	**22**	**26**	22	24
10	19	17	**24**	22	21	22	**30**	27	28

Results

Table 3. MRR results on *TF-IDF* dataset. Note that *NCS* + *NQE* outperforms *NCS* on short queries of length 1 and 2.

Query Length	Mean Reciprocal Rank					
	NCS	NCS + FIM	NCS + NQE	BM25	BM25 + FIM	BM25 + NQE
1	0.092	0.109	**0.284**	0.060	0.045	**0.219**
2	0.416	0.428	**0.543**	0.276	0.193	**0.390**
3	**0.672**	0.547	0.574	**0.528**	0.356	0.424
4	**0.807**	0.706	0.650	**0.657**	0.494	0.542
5	**0.852**	0.727	0.679	**0.649**	0.491	0.531
6	**0.951**	0.839	0.812	**0.729**	0.574	0.605

Results

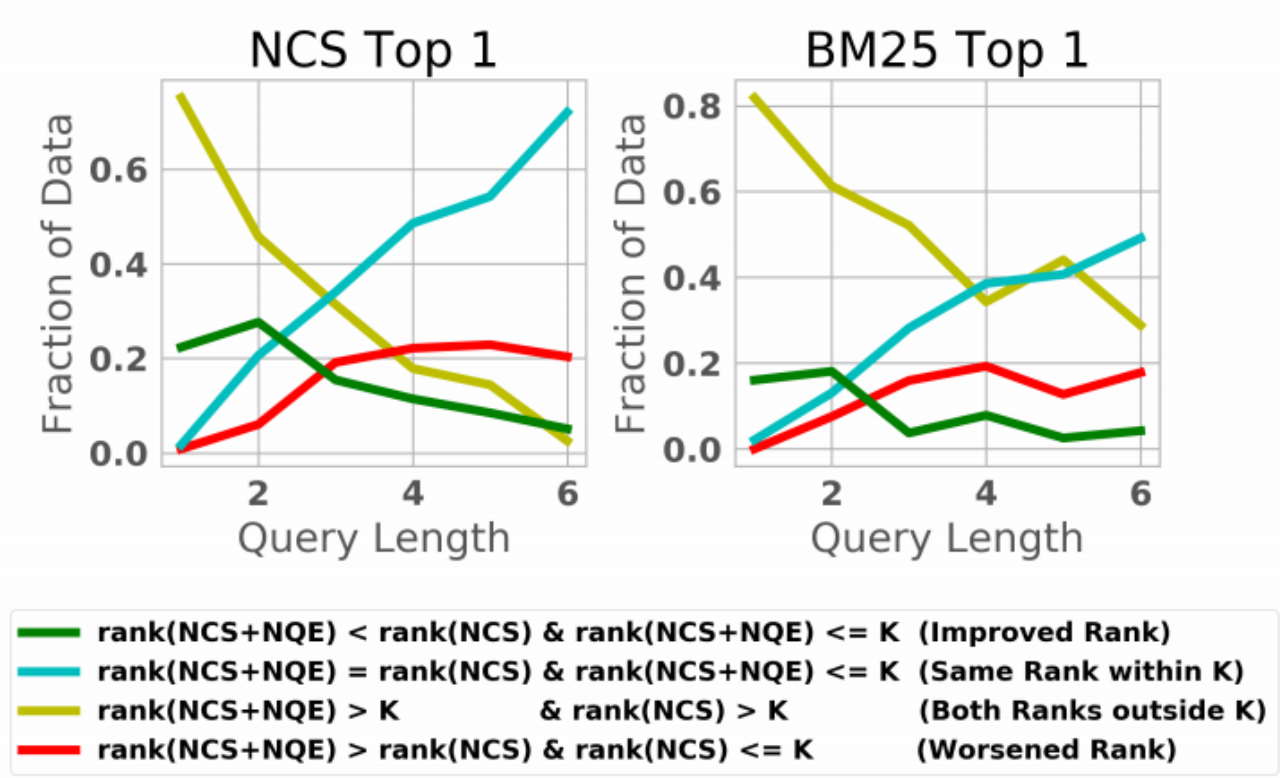


Figure 6. Top 1 ranking changes between *NCS+NQE* vs *NCS*.

Results

Table 4. MRR results on *Manual* dataset. Note similar trends to Table 3.

Query Length	Mean Reciprocal Rank					
	NCS	NCS + FIM	NCS + NQE	BM25	BM25 + FIM	BM25 + NQE
1	0.040	0.080	**0.178**	0.035	0.049	**0.139**
2	0.319	0.292	**0.352**	0.272	0.258	**0.310**
3	**0.545**	0.381	0.456	**0.440**	0.310	0.396
4	**0.706**	0.438	0.560	**0.573**	0.364	0.492
5	**0.782**	0.430	0.609	**0.690**	0.378	0.547
6	**0.814**	0.481	0.626	**0.721**	0.452	0.589

Results

RQ2: How does the quality of NQE sequence prediction affect the end-result?

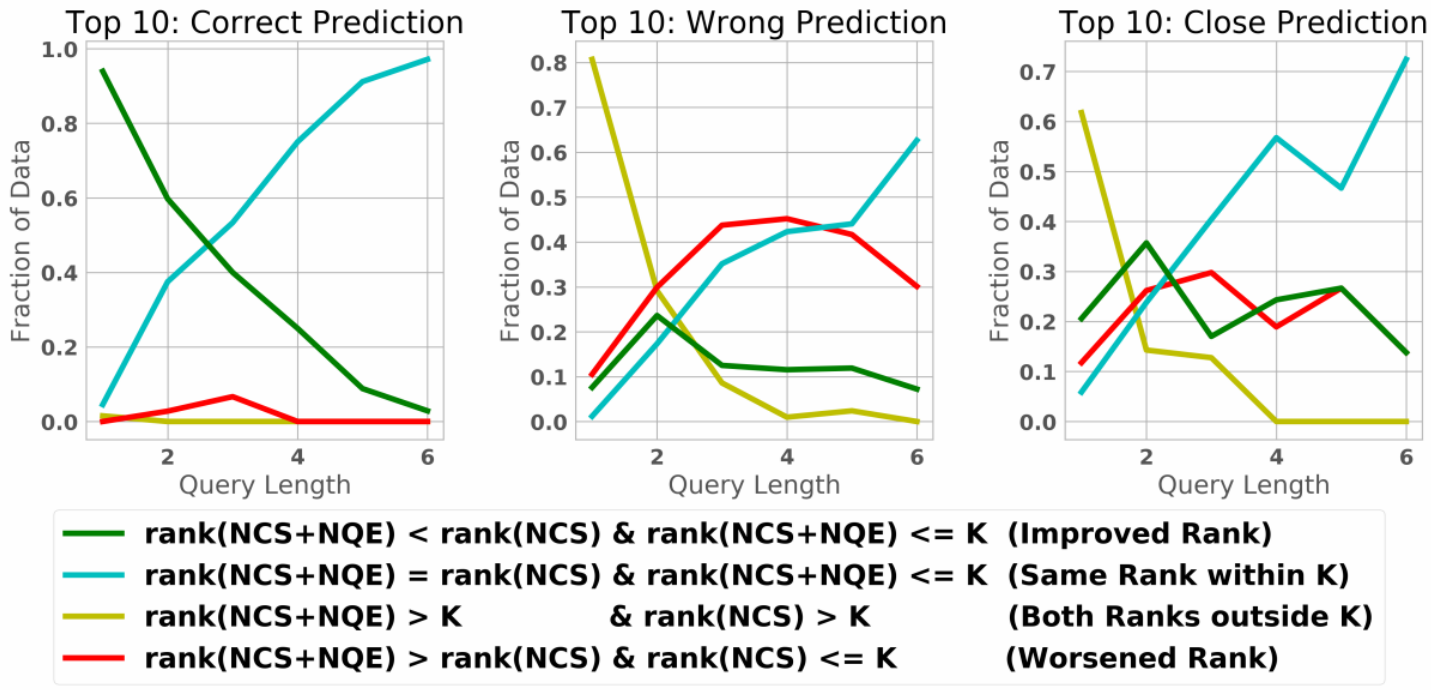


Figure 7. Top 10 ranking changes between *NQE + NCS* vs *NCS* when the prediction is correct, wrong, or close on the *TF-IDF* dataset.

Results

RQ3: Do these findings generalize to other search

CONTENTS

Future Work

1、NQE的一个修改是在token的子集上进行序列预测

优点：1）较小的解码器词汇集大小

2）处理包含公共token的稀有方法名称的能力

2、将NQE与集合扩展方法进行对比