Distractor Generation for Multiple Choice Questions Using Learning to Rank

Chen Liang[1], Xiao Yang[2], Neisarg Dave[1],
Drew Wham[3], Bart Pursel[3], C. Lee Giles[1,2]
[1]Information Sciences and Technology [2]Computer Science and Engineering [3]Teaching and Learning with Technology
Pennsylvania State University, University Park, PA, USA

PennState

Background

- Distractor Generation (DG): generate distractors given the question sentence (stem) and the correct answer (key)

Input

Q: A compound which is found in all living cells and play a key role in energy transformation is _____.
A: ATP

Output

ADP
Chlorophyll
Granum

Motivation

- DG is a crucial step for MCQ generation.
- Previous DG methods were mostly based on semantic similarities between the key and the candidate distractor
  - WordNet synonyms [Mitkov, BEA'03]
  - Phonetic and morphological similarity [Pino, SLATE'09]
  - Embedding-based similarity [Kumar, BEA'15; Jiang, BEA'17]
  - Co-occurrence likelihoods [Hill, BEA'16]

Unsupervised, heuristic-based

Learning to Rank for DG

Problem. Given a candidate distractor set $\mathcal{D}$ and a MCQ dataset $\mathcal{M} = \{(q_i, a_i, \{d_{i1}, ..., d_{ik}\})\}_{i=1}^{N}$, where q_i is the question stem, a_i is the key, $D_i = \{d_{i1}...d_{ik}\} \subseteq \mathcal{D}$ are the distractors associated with q_i and a_i, find a point-wise ranking function r: $(q_i, a_i, d) \rightarrow [0, 1]$ for $d \in \mathcal{D}$, such that distractors in D_i are ranked higher than those in $\mathcal{D} - D_i$.

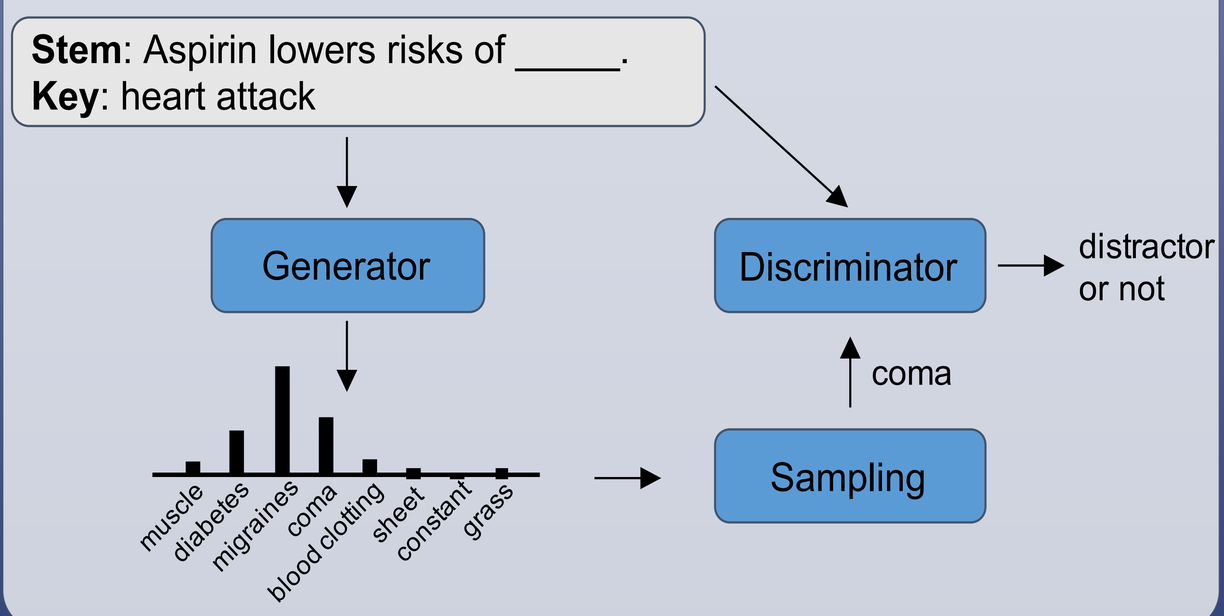
Feature-based Models

- Models: (i) Logistic Regression; (ii) Random Forest; (iii) LambdaMART

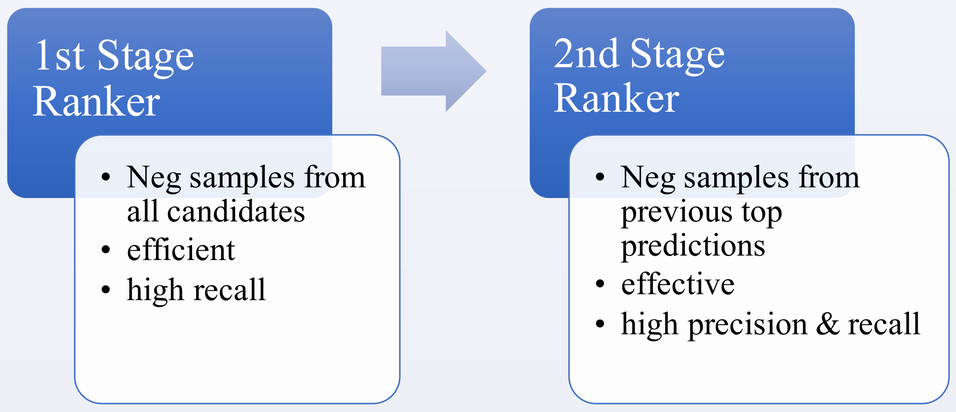
Feature	Description
Emb Sim	Embedding similarity between q and d and the similarity between a and d
POS Sim	Jaccard similarity between a and d's POS tags
ED	Edit distance between a and d
Token Sim	Jaccard similarities between q and d's tokens, a and d's tokens, and q and a's tokens.
Length	a and d's character and token lengths and the difference of lengths.
Suffix	The absolute and relative length of a and d's longest common suffix.
Freq	Average word frequency in a and d.
Single	Singular/plural consistency of a and d.
Num	Whether numbers appear in a and d.
Wiki Sim	Wiki embedding similarity.

NN-based Models

- An adversarial training framework based on IRGAN [Wang, SIGIR'17]



Cascaded Learning Framework



Experiments

Dataset

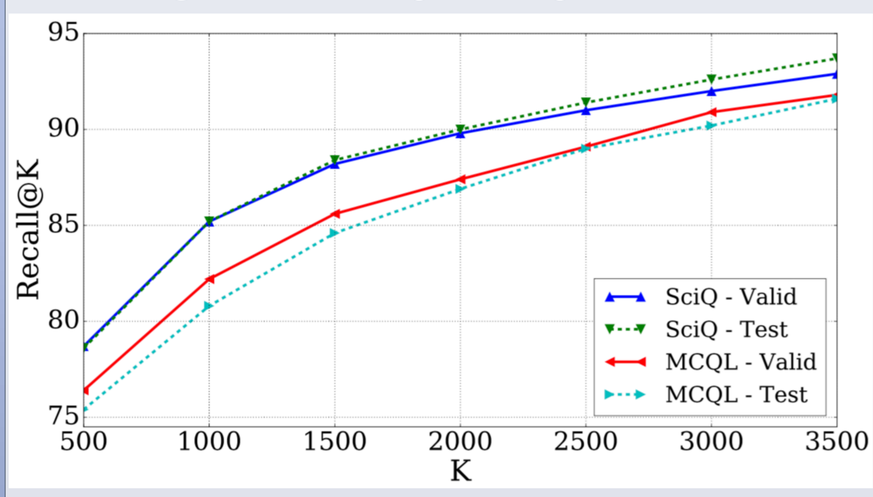
Dataset	$\lvert\mathcal{D}\rvert$	# MCQs	# Train	# Valid	# Test	Avg. # Dis
SciQ	22379	13679	11679	1000	1000	3
MCQL	16446	7116	5999	554	563	2.91

SciQ: [Welbl, 2017];
MCQL: collected by us.

Dataset available:
https://github.com/harrylclc/LTR-DG

Experiment Results

First Stage Ranker: Logistic Regression



Thresholds:
SciQ: K=2000
MCQL: K=2500

Figure: Recall@K for the first stage ranker

Ranking Results (%) for DG

1st Stage Ranker	2nd Stage Ranker	R@10	P@1	P@3	MAP @10	NDCG @10	MRR
LR	PMI	11.0	2.1	3.1	3.6	6.8	8.8
	ED	14.3	12.6	9.2	8.7	12.5	18.9
	Emb Sim	19.3	9.3	9.0	9.6	14.2	17.5
	LR	29.7	14.8	14.1	14.7	22.1	27.6
	RF	**44.1**	36.8	**27.0**	**28.4**	**38.0**	49.2
	LM	43.3	**37.2**	26.4	28.0	37.5	**49.3**
	NN	24.6	11.7	11.7	11.6	23.1	25.7
RF	—	41.4	31.2	23.7	25.0	34.4	44.0
LM	—	39.1	26.5	22.6	22.9	31.8	40.4

(a) SciQ

1st Stage Ranker	2nd Stage Ranker	R@10	P@1	P@3	MAP @10	NDCG @10	MRR
LR	PMI	20.7	5.9	6.8	7.8	13.5	16.2
	ED	32.1	34.6	23.6	23.7	30.5	42.8
	Emb Sim	32.1	25.6	18.4	20.4	26.9	33.9
	LR	42.9	29.3	24.5	26.6	35.1	42.2
	RF	48.4	**45.5**	**32.7**	**35.4**	**43.8**	**54.8**
	LM	**49.4**	42.8	31.5	34.5	43.4	53.6
	NN	36.5	22.9	22.5	22.7	34.6	36.7
RF	—	48.0	40.9	30.4	33.6	42.0	51.1
LM	—	46.7	42.5	30.6	33.0	41.6	52.7

(b) MCQL

Findings:
1. Ranking models > unsupervised baselines
2. Ensemble methods (RF and LM) outperform others
3. NN-based models < feature-based models

Future Work

- Deploy the proposed methods to actual MCQ systems
- Evaluate DG methods using user feedback

Acknowledgements: We gratefully acknowledge partial support from the Pennsylvania State University Center for Online Innovation in Learning.