

ID: 252

Candidate Registration C1 DATE FILED PDC JUN 19 2008

Candidate Name: Anna M. Rivers

Office: State Senator

City: La Center

County: Clark

Phone Number: 360-931-3403

ID: 454

Candidate Registration C1 DATE FILED PDC JUN 19 2008

Candidate Name: Clifford M. Greene

Office: State Representative

City: Federal Way

County: King

Phone Number: 253-838-1838

ID: 1298

Candidate Registration C1 DATE FILED PDC JUN 19 2008

Candidate Name: Lucius Dave

Office: Thurston Commissioner

City: Lacey

County: Thurston

Phone Number: 360-459-4986

ID: 8358

Candidate Registration C1 DATE FILED PDC JUN 19 2008

Candidate Name: Clifford M. Greene

Office: State Representative

City: Federal Way

County: King

Phone Number: 253-838-1838

ID: 10901

Candidate Registration C1 DATE FILED PDC JUN 19 2008

Candidate Name: Anna M. Rivers

Office: State Senator

City: La Center

County: Clark

Phone Number: 360-931-3403

Question: In which years did Anna M. Rivers run for the State senator office?
Answer: [2016, 2020]
Evidences: [454, 10901]

Question: Have there been any representative of the Green Party in Seattle?
Answer: Yes
Evidences: [4000]

Overview

- Motivation: Lack of high-level purpose of DAR and dataset collection tasks.
- Images are the US Candidate Registration form, sourced from the Open data portal of the Public Disclosure Commission.
- 20 Questions posed over 14,362 document images.
- Methods must provide the answer to the question but also the document IDs of the documents used to obtain the answer (positive evidence).
- New metric based on ANLS to assess methods answering performance on a set of answers regardless of their order.
- New baselines are proposed and evaluated to showcase the performance of different approaches on this task.

Metric: Average Normalized Levenshtein Similarity for Lists (ANLSL)

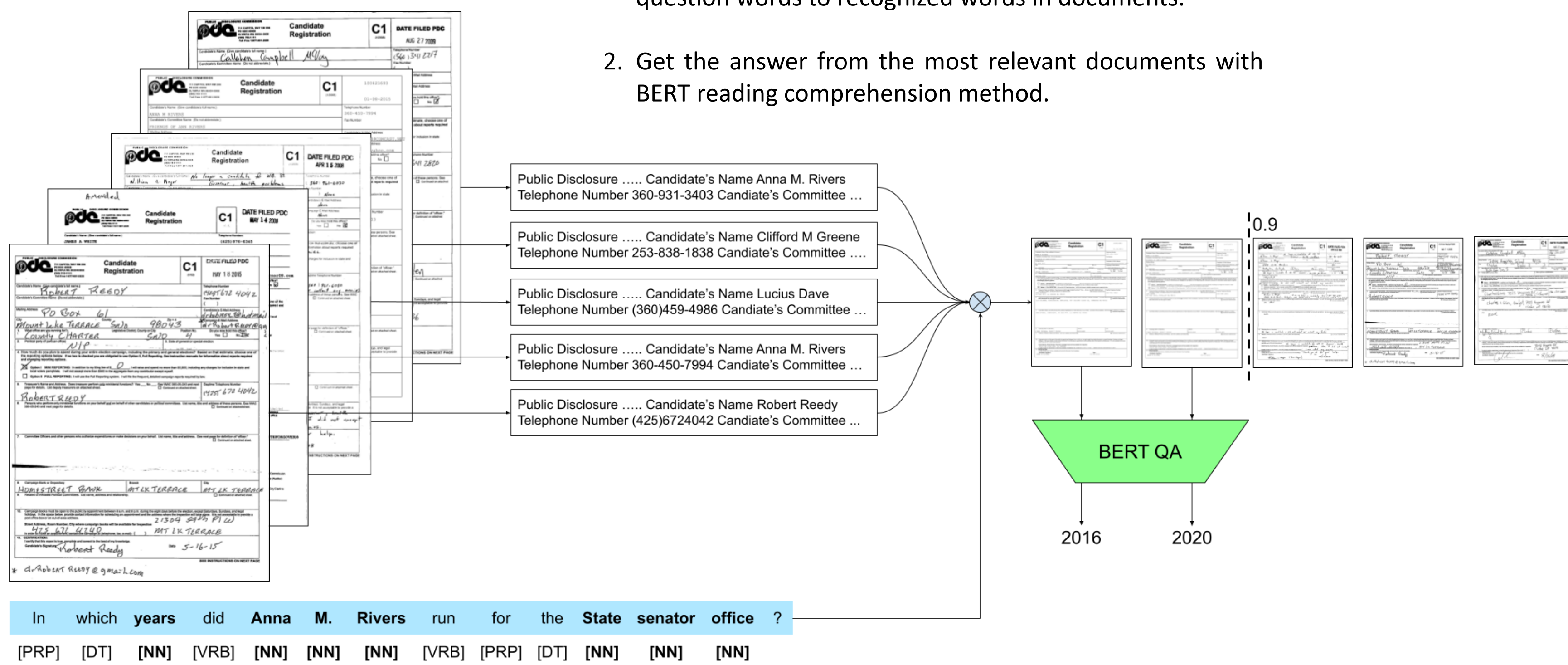
- Based on string edit distance.
- Smoothly penalizes OCR recognition errors while evaluating method's reasoning capabilities on itemized answers.
- Hungarian Matching (Ψ) to find closest predictions-ground truth pairs (U).

$$U = \Psi(NLS(G, P)) \quad ANLSL = \frac{1}{\max(M, N)} \sum_{z=1}^K NLS(u_z)$$

Baselines

Text Spotting + BERT

- Rank documents according to edit distance from relevant question words to recognized words in documents.
- Get the answer from the most relevant documents with BERT reading comprehension method.



Database

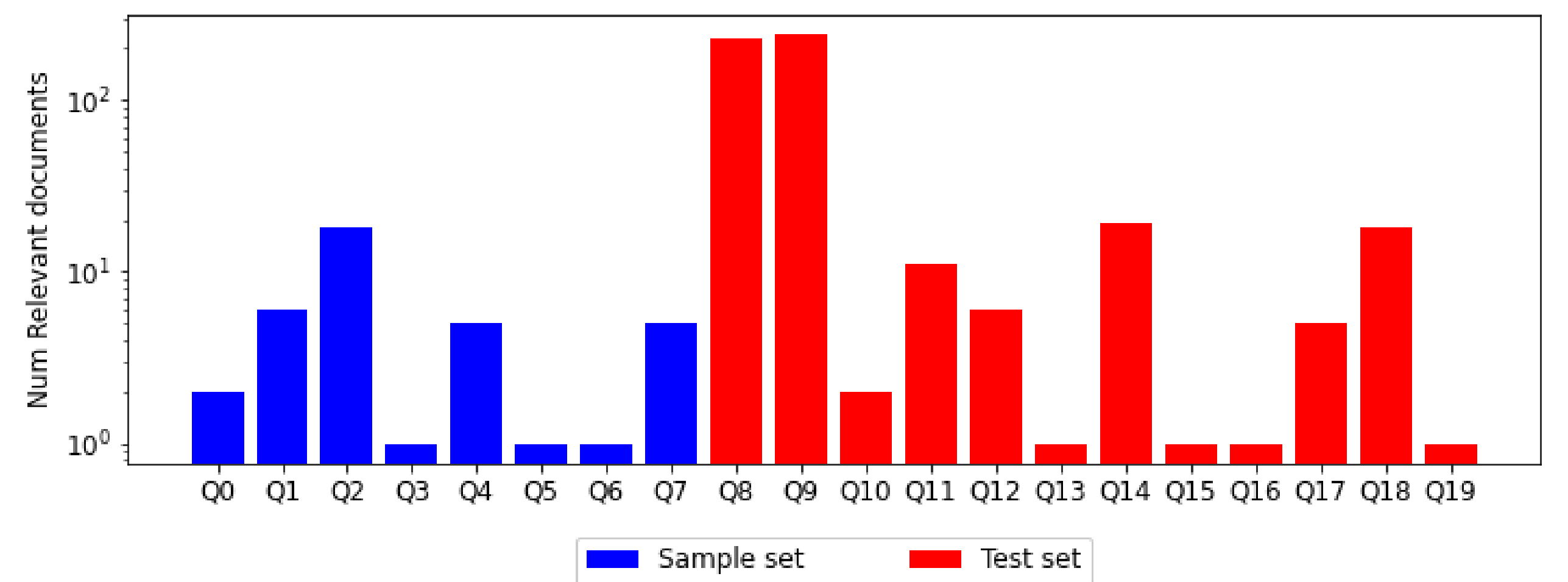
- Extract key-value relations on documents and dump it into a database-like structure.
- Manually parse the question into Structured Query Language to get both relevant documents and the answer.

Name	Office	City	County	Phone Number
Anna M. Rivers	State Senator	La Center	Clark	360-931-3403
Clifford M. Greene	State Represent.	Federal Way	King	253-838-1838
Lucius Dave	Thurston Comm.	Lacey	Thurston	(360)459-4986
Anna M. Rivers	State Senator	La Center	Clark	360-450-7994
Robert Reedy	County Charter	Mount Lake Terrace	Snohomish	(425)6724042

In which years did Anna M. Rivers run for the State senator office? — SELECT election_date, year WHERE Name="Anna M. Rivers" and Office="State senator"

DocCVQA

Field	Type	# Values	# Unique Values
Candidate name	Text	14362	9309
Party	Text	14161	10
Office	Text	14362	43
Candidate city	Text	14361	476
Candidate county	Text	14343	39
Election date	Date	14362	27
Reporting option	Checkbox	14357	2
Treasurer name	Text	14362	10197



Results

Retrieval method	Answering Method	MAP	ANLSL
Text spotting	BERT	72.84	0.4513
Database	BERT	71.06	0.5411
Database	Database	71.06	0.7068
GT	BERT	100.00	0.5818
GT	Database	100.00	0.8473

