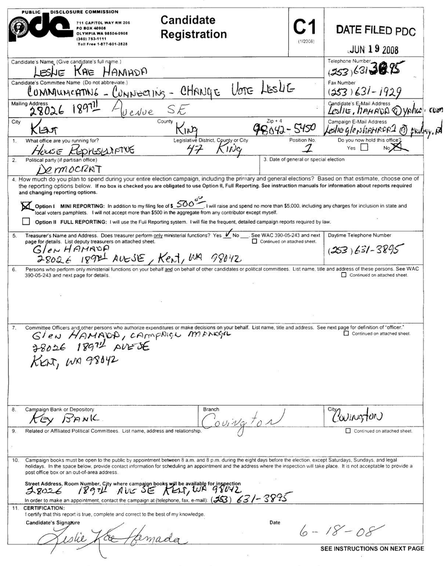
# Document Collection Visual Question Answering

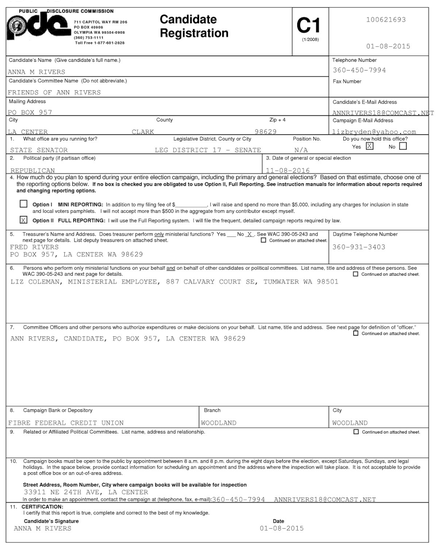
Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny

Computer Vision Center, UAB, Spain

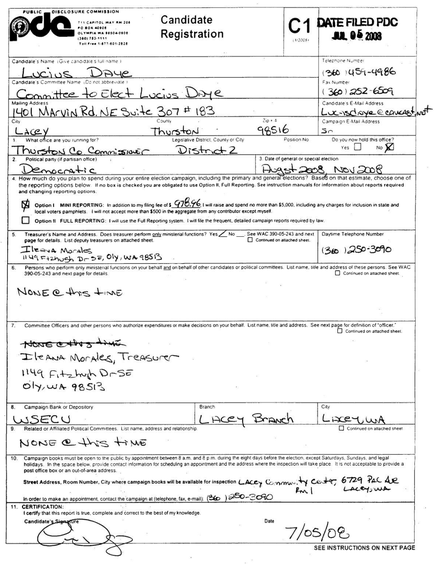
ID: 252



ID: 454

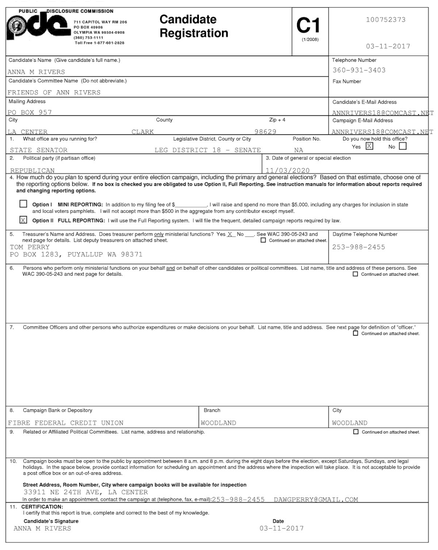


ID: 1298



ID: 8358

ID: 10901



**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.

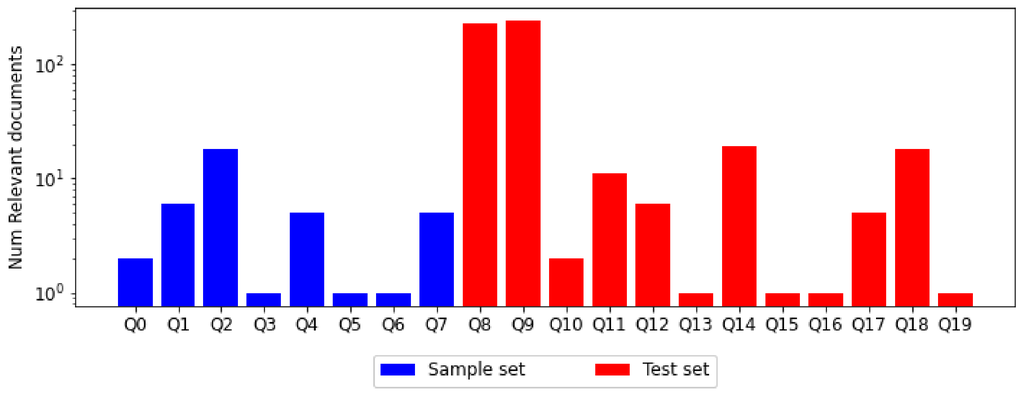
## DocCVQA

| 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 |

## Metric: Average Normalized Levenshtein Similarity for Lists (ANLSL)

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

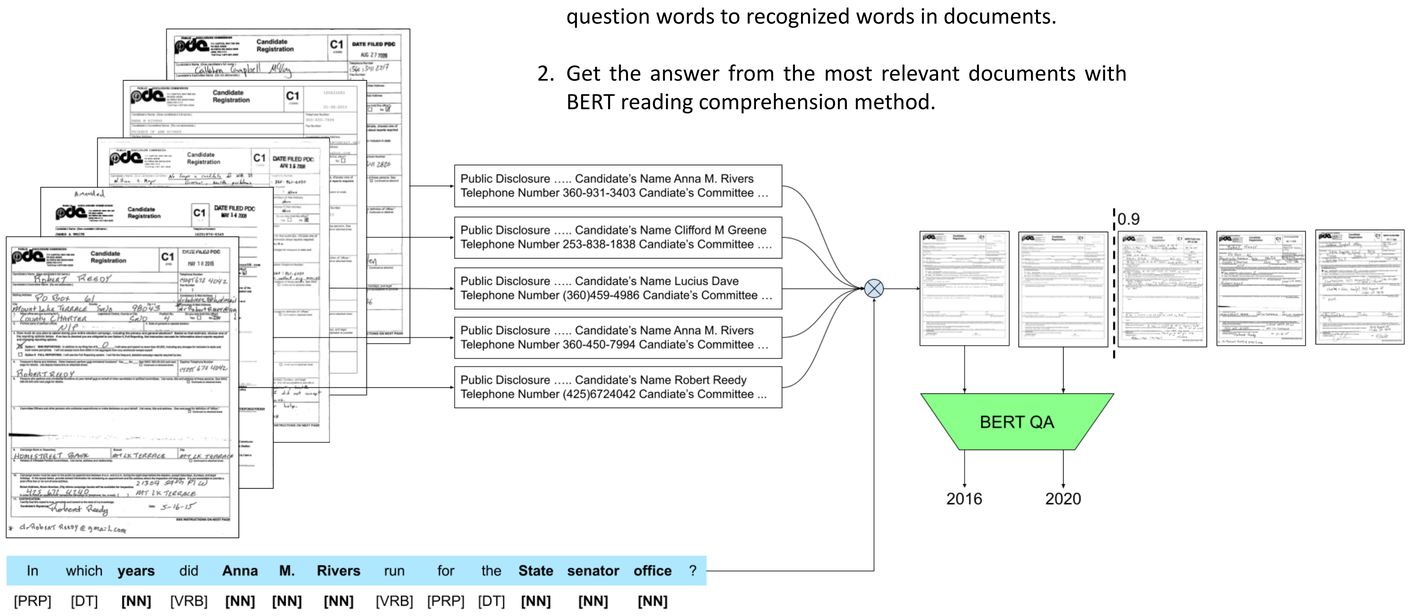
$$U = \Psi\big(NLS(G,P)\big) \qquad ANLSL = \frac{1}{\max(M,N)} \sum_{z=1}^{K} NLS(u_z)$$




## Baselines

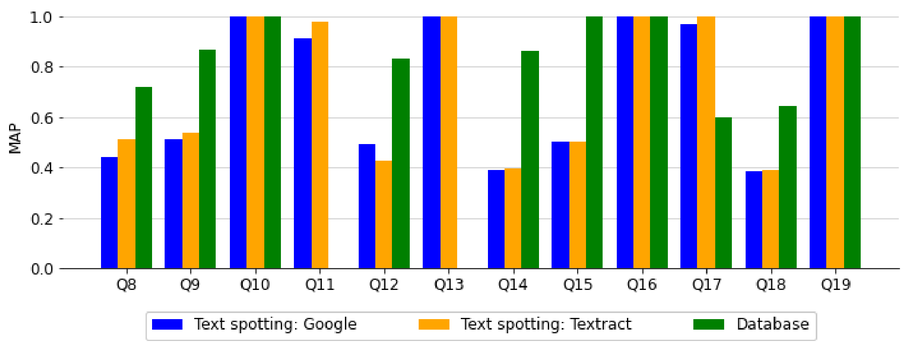
### Text Spotting + BERT

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




### 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 |




### Database

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