Rich Context Competition

Team KAIST

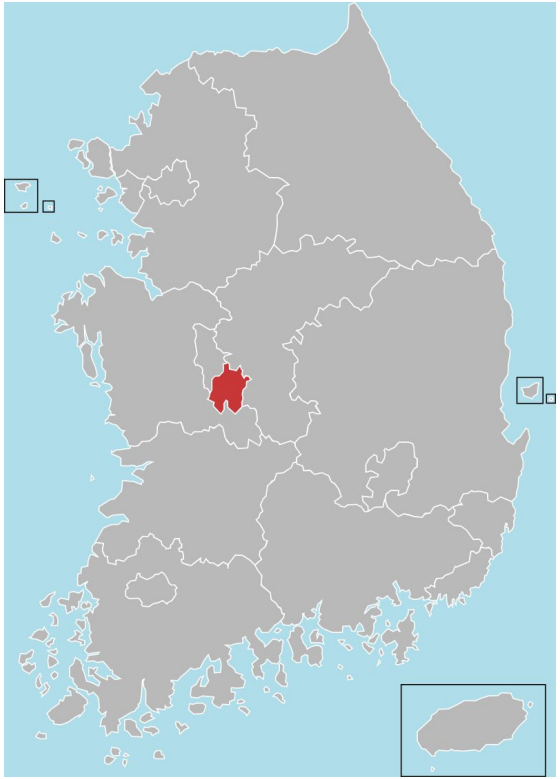
RCC Workshop 15th Feb 2019

KAIST

Introduction

- Giwon Hong from Korea
- Son Minh Cao from Vietnam
- Haritz Puerto-San-Roman from Spain

- Master's students at KAIST, Daejeon, South Korea
- IR&NLP Lab, School of Computing



Introduction

- Task definition:
  - Obtain dataset names, research fields and methods from a collection of scientific publications
- Our approach:
  - Reading Comprehension QA, entity typing
  - TF-IDF similarity
  - Named-Entity Recognition

Datasets

- Our approach to retrieve datasets
  - **Reading comprehension** (RC) model
  - With **our own generated queries**
  - And filtering by **entity types**

Datasets - RC

- Reading comprehension (RC) QA models
  - Neural network models to find answers for given queries and texts
  - Answers are usually specific spans from texts

Datasets - RC

- Reading comprehension (RC) QA models
  - Neural network models to find answers for given queries and texts
  - Answers are usually specific spans from texts

Query

When is the april fools day?

Text

April Fools' Day (sometimes called All Fools' Day) is celebrated every year on **April 1** by playing practical jokes and spreading hoaxes.

Datasets - RC

- Reading comprehension (RC) QA models
  - Neural network models to find answers for given queries and texts
  - Answers are usually specific spans from texts

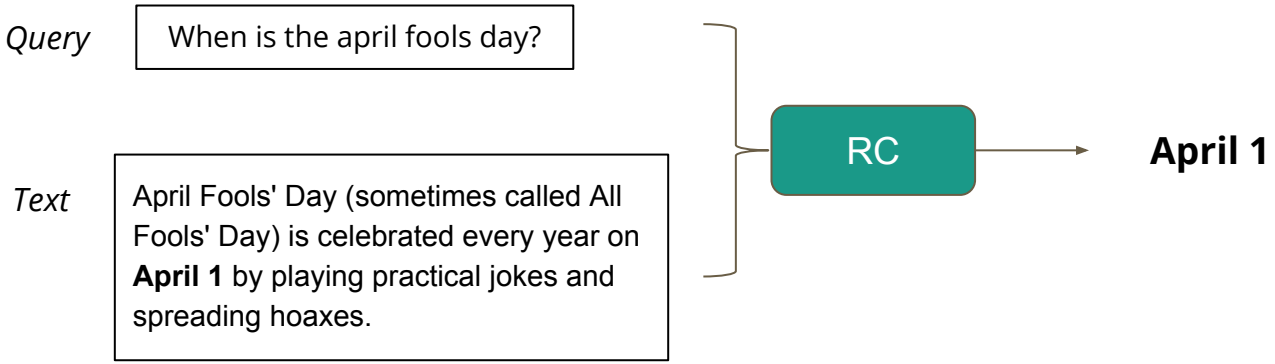
Query When is the april fools day?

Text April Fools' Day (sometimes called All Fools' Day) is celebrated every year on **April 1** by playing practical jokes and spreading hoaxes.

RC

Datasets - RC

- Reading comprehension (RC) QA models
  - Neural network models to find answers for given queries and texts
  - Answers are usually specific spans from texts



Datasets - RC

- In Rich Context Competition
  - Text: Publications in social science
  - Answer: dataset mentions in publications
  - Which RC model?
  - Which query?

Datasets - RC

- Document QA model
  - Clark et al., 2017
  - **RC model** with **paragraph selection**

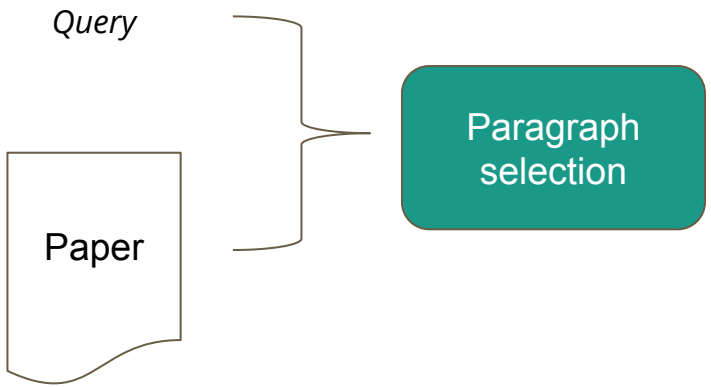
Datasets - RC

- Document QA model
  - Clark et al., 2017
  - **RC model** with **paragraph selection**

Query

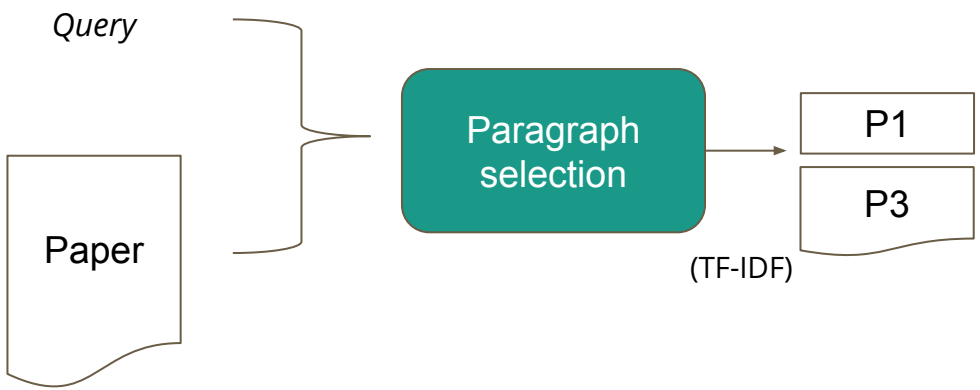
Datasets - RC

- Document QA model
  - Clark et al., 2017
  - **RC model** with **paragraph selection**



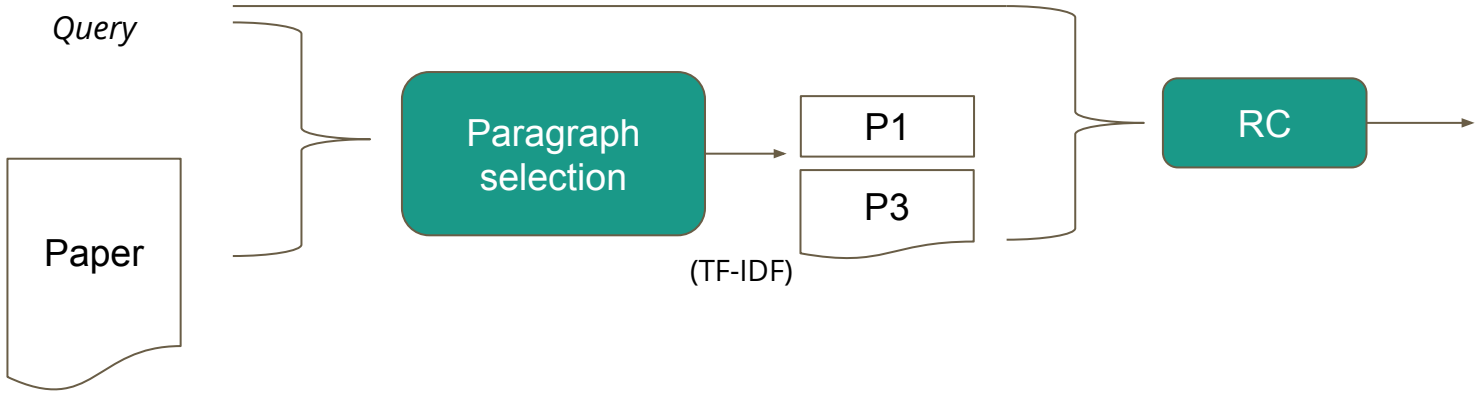
Datasets - RC

- Document QA model
  - Clark et al., 2017
  - **RC model** with **paragraph selection**



Datasets - RC

- Document QA model
  - Clark et al., 2017
  - **RC model** with **paragraph selection**



Datasets - RC

- Why Document QA?
  - The mentions tend to **cluster in certain parts** of the publications

    ⇒ Finding that certain parts (paragraphs)

    ⇒ **Paragraph selection** in Document QA

Datasets - RC

- Query?
  - We need queries to retrieve dataset mentions
  - However, it is **difficult** to find **general queries** since datasets appear in various form
  - Examples:

Datasets - RC

- Query?
  - We need queries to retrieve dataset mentions
  - However, it is **difficult** to find **general queries** since datasets appear in various form
  - Examples:

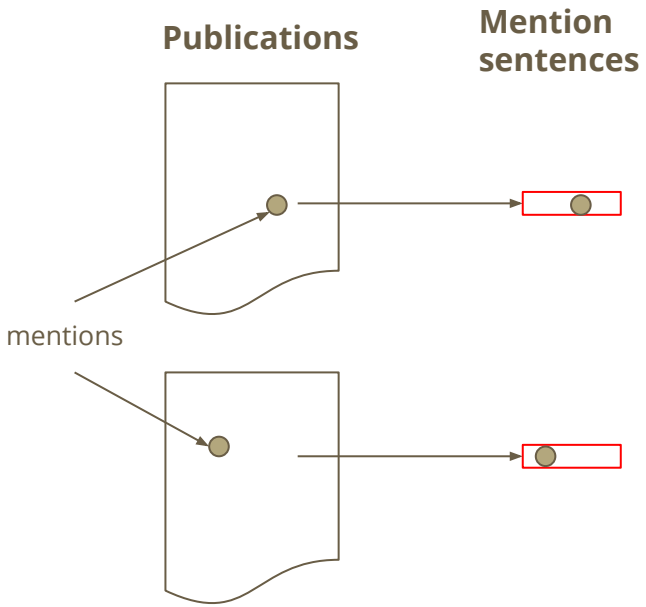
ANES 1952 Time Series Study		What study?
Survey of State Court Criminal Appeals, 2010	⟷	What survey?
National Material Capabilities dataset		What dataset?

Datasets - Query

- We focus on
  - Making **many queries**, instead of one general query
  - Queries with enough **discriminative power** to retrieve dataset mentions
  - Generating important **query terms**

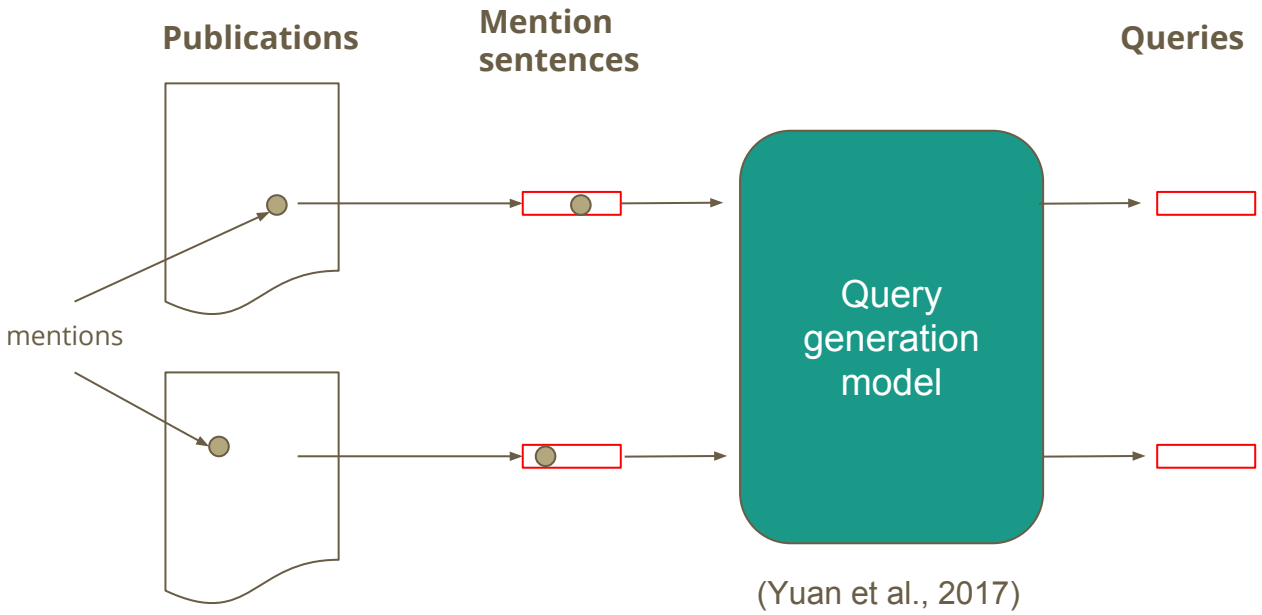
Datasets - Query terms

1) Extract sentences that contain dataset mentions (training set)



Datasets - Query terms

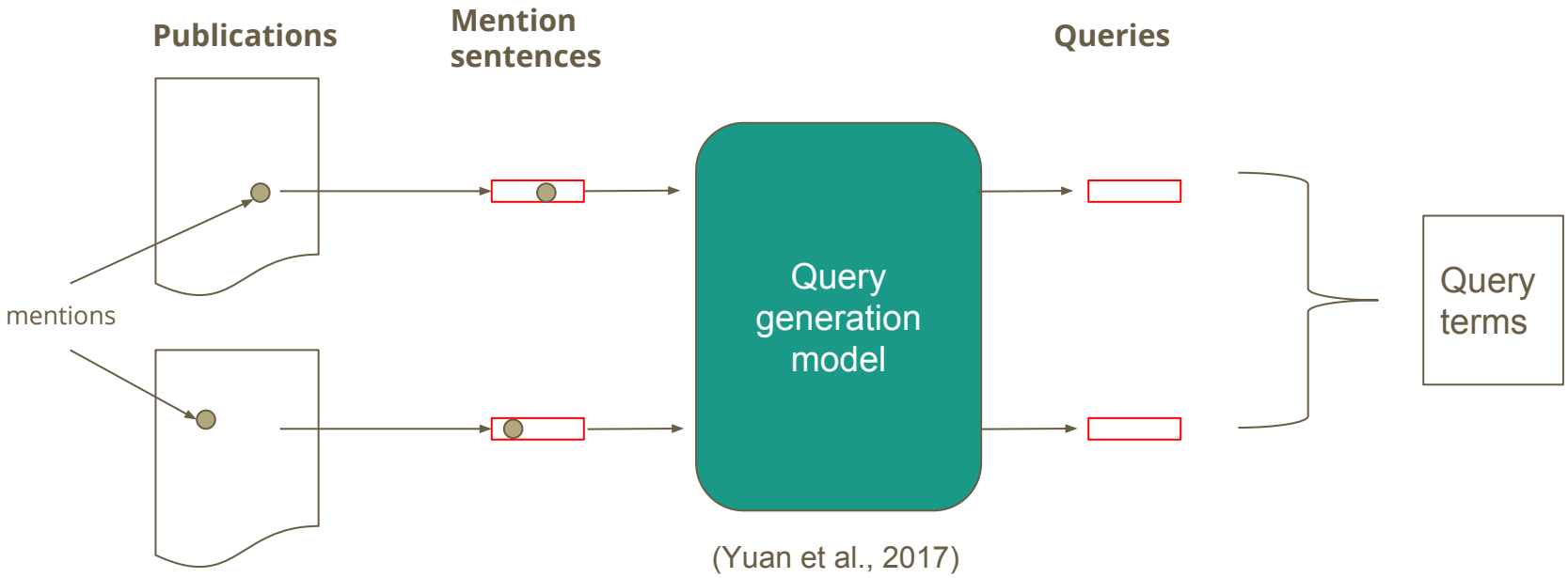
2) Generate queries that can find mention from sentences



(Yuan et al., 2017)

Datasets - Query terms

3) Find query terms from queries



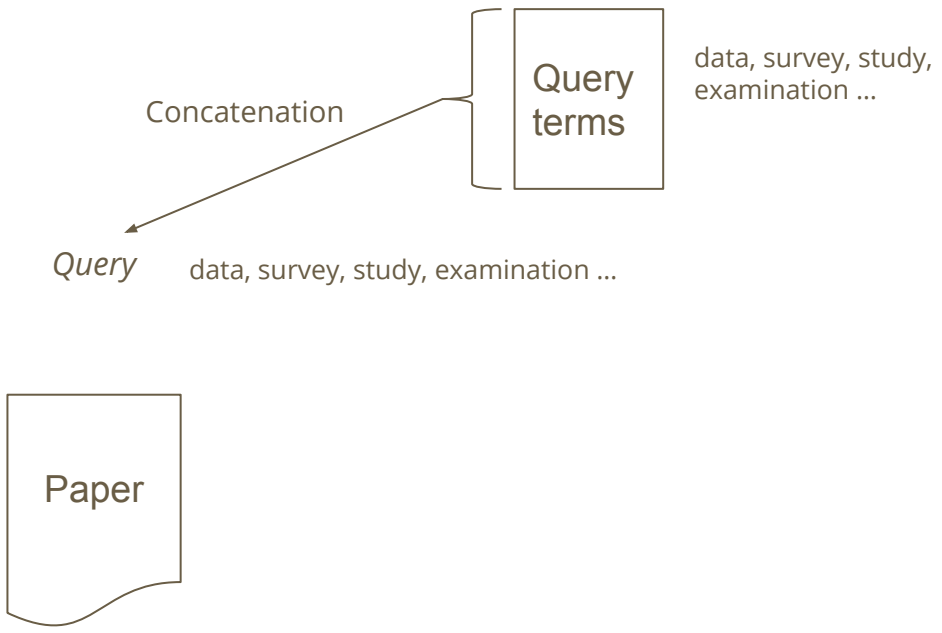
(Yuan et al., 2017)

Datasets - Query + RC

- We use these terms to generate queries for each paragraph on the fly

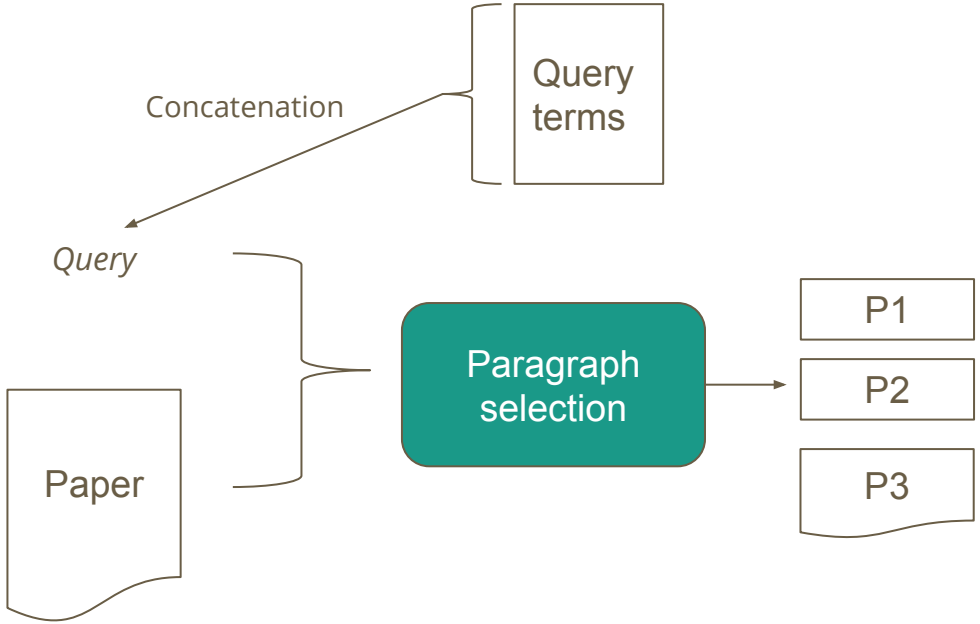
Datasets - Query + RC

1) Generate a general query by concatenating query terms for **paragraph selection**



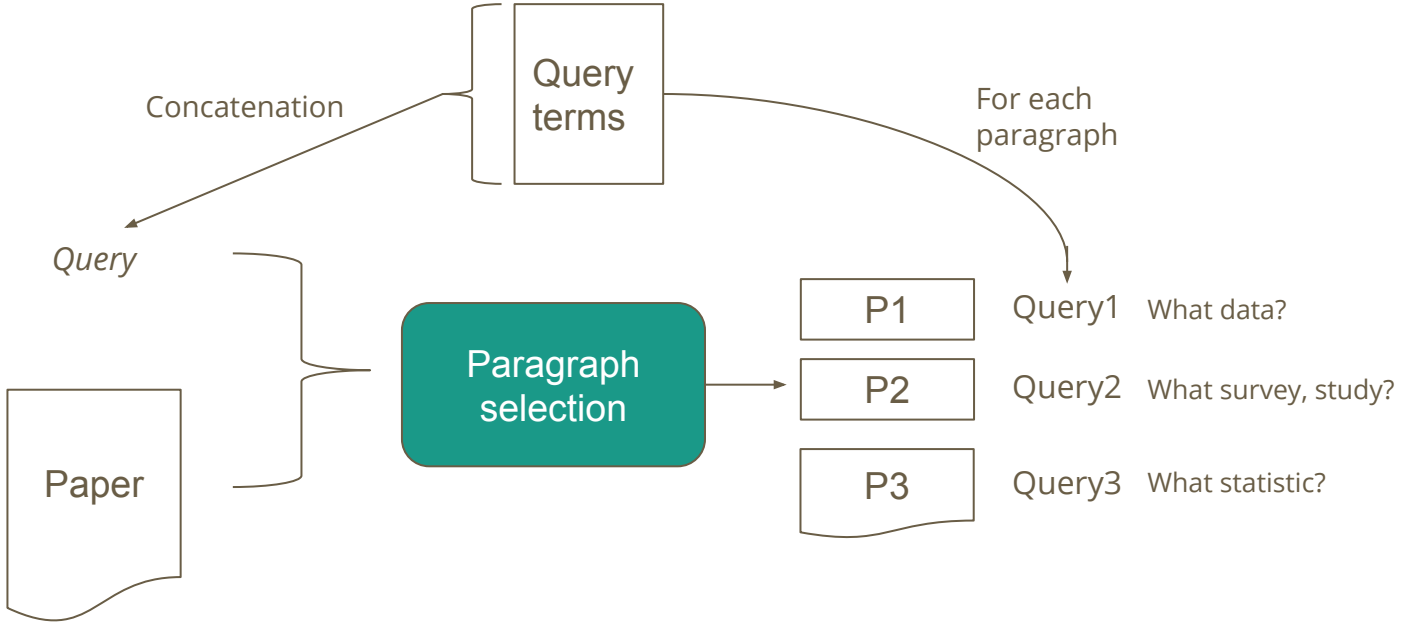
Datasets - Query + RC

2) Paragraph selection



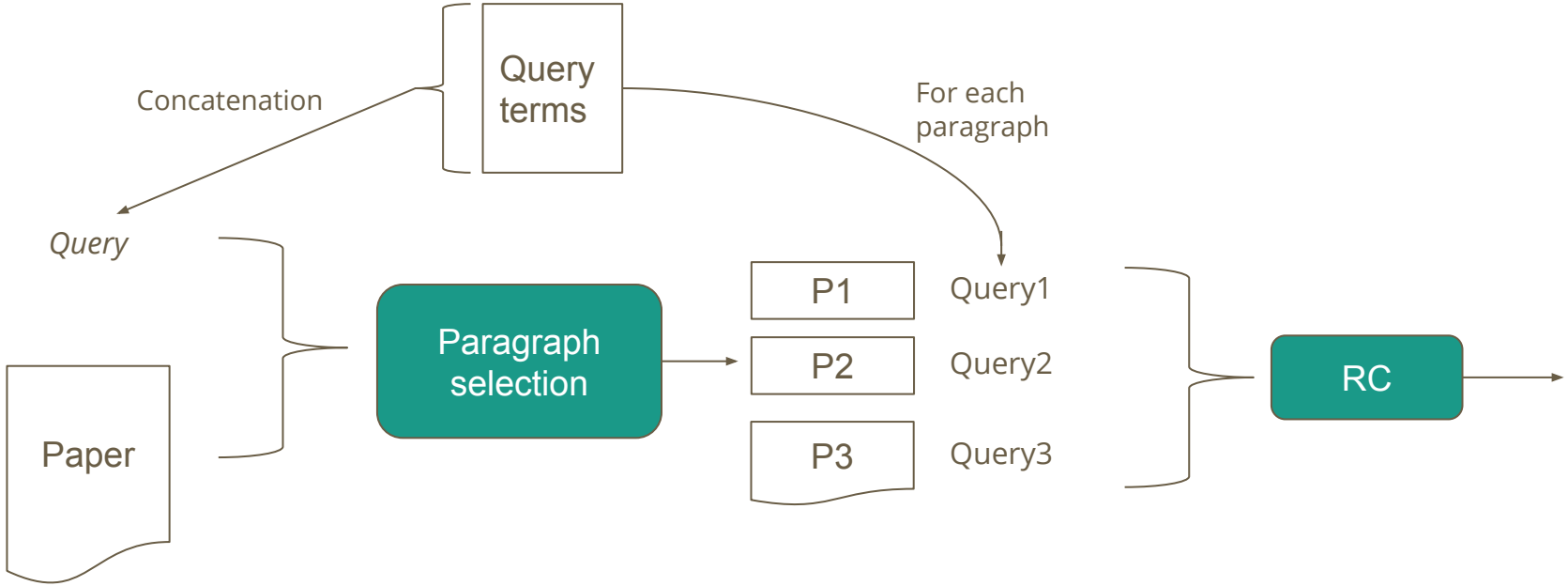
Datasets - Query + RC

3) Generate queries for each paragraph with query terms



Datasets - Query + RC

4) Input to RC model



Datasets - Results from DocQA

- DocQA is able to retrieve right answers (datasets)

Datasets - Results from DocQA

- DocQA is able to retrieve right answers (datasets)
- However, it has a lot of noise

1134.txt

- British Psychiatric Morbidity Survey
- National Comorbidity Survey
- The National Comorbidity Survey
- NCS
- **Table 1**
- **psychosis**

153.txt

- financial services FDI data
- **empirical**
- **Deutsche Bundesbank (the German central bank)**
- Micro Database Direct Investment
- **/ / go.worldbank.org / SNUSW978P0"**
- **mixed logit model**

143.txt

- **Empirical**
- ITS data
- **collective reports**
- **transactions below the reporting limit of e12,500**
- **Section 4**
- **4 2.1 Micro Data**

Datasets - Results from DocQA

- DocQA is able to retrieve right answers (datasets)
- However, it has a lot of noise
- We need to remove that noise

Datasets - Results from DocQA

- DocQA is able to retrieve right answers (datasets)
- However, it has a lot of noise
- We need to remove that noise
  - Filtering by **Entity types**
  - Dataset names share very similar entity types (organization, agency, etc)

Datasets - Ultra Fine Entity Typing

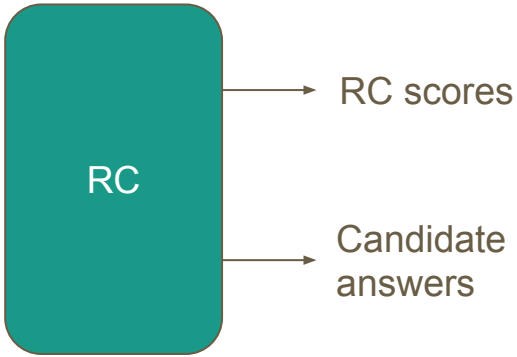
"We use the ***Deutsche Bundesbank balance of payments statistics*** as our main source of data" → **Organization**, bank

- Can predict 10k different entity types
- Choi et al., 2018

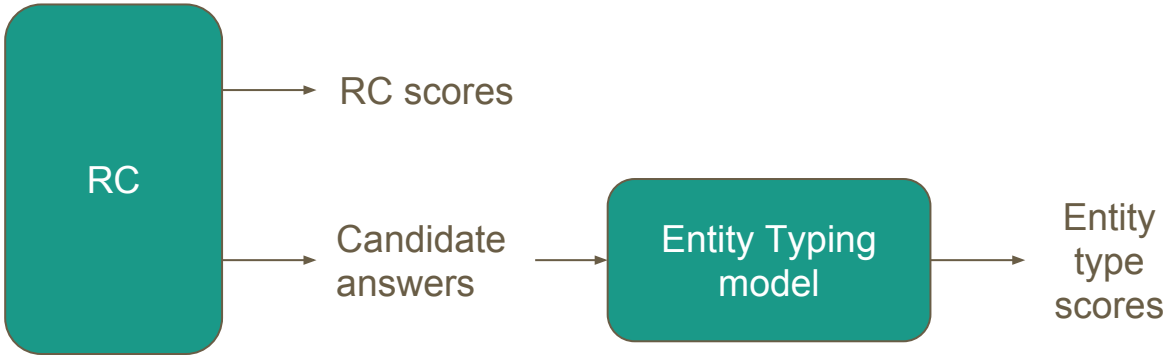
Datasets - Answer Classifier

- Thanks to the **entity types**, we can **filter out** candidate answers from the RC model
- We trained a **NN** that classifies right answers using:
  - **Scores** from the **RC** model
  - **Scores** from the **entity typing** model

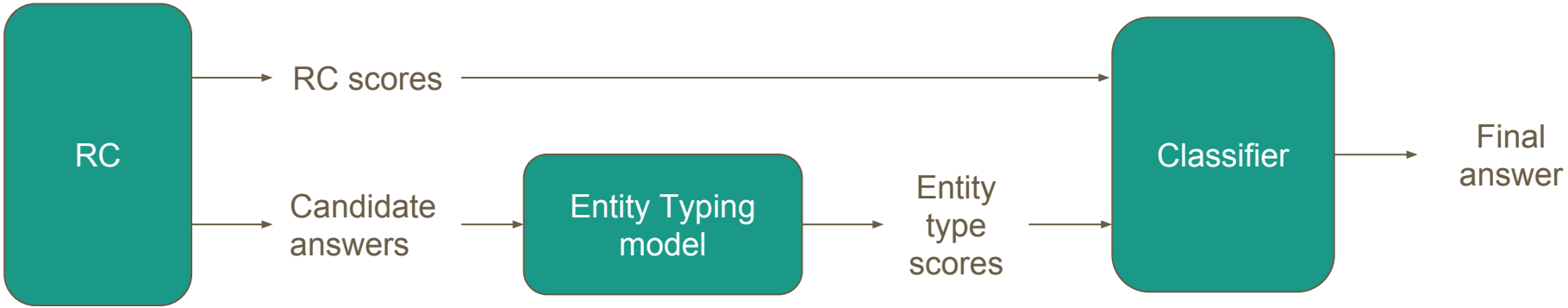
Datasets - Overall Architecture



Datasets - Overall Architecture



Datasets - Overall Architecture



Datasets - Analysis

- Effects of the **query generation** and the **entity typing**
  a. Document QA
  b. Document QA + query generation
  c. Document QA + query generation + entity typing

Datasets - Analysis

- Document QA
  - 260 answers from 100 publications

1134.txt

- **National Comorbidity Survey**

153.txt

- None

143.txt

- **MiDi**
- **the Balance of Payments Statistics**

Datasets - Analysis

- Document QA + query generation
  - 2,000 answers from 100 publications

1134.txt

- **British Psychiatric Morbidity Survey**
- **National Comorbidity Survey**
- **The National Comorbidity Survey**
- **NCS**
- Table 1
- psychosis

153.txt

- **financial services FDI data**
- empirical
- Deutsche Bundesbank (the German central bank)
- **Micro Database Direct Investment**
- / / go.worldbank.org / SNUSW978P0"
- mixed logit model

⋮

143.txt

- Empirical
- **ITS data**
- collective reports
- transactions below the reporting limit of e12,500
- Section 4
- 4 2.1 Micro Data

⋮

Datasets - Analysis

- Document QA + query generation + entity typing
  - 526 answers from 100 publications

1134.txt

- **British Psychiatric Morbidity Survey**
- **National Comorbidity Survey**
- **The National Comorbidity Survey**
- **NCS**

153.txt

- **Micro Database Direct Investment**

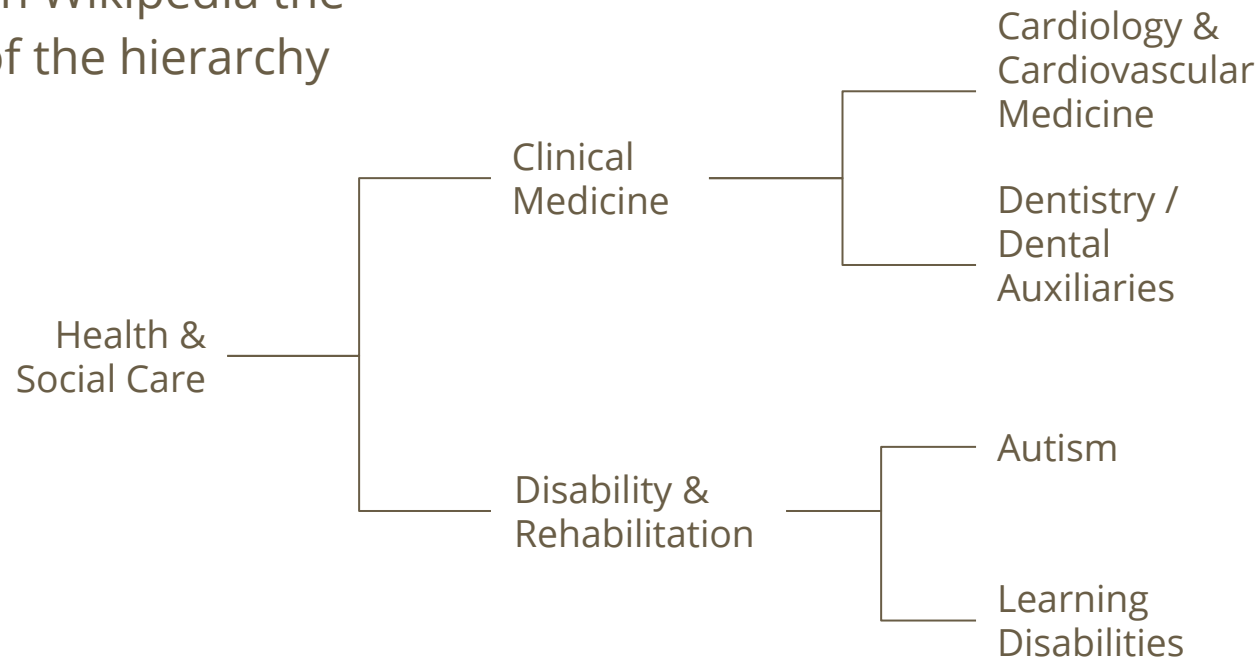
143.txt

- **ITS data**
- 4 2.1 Micro Data
- determinants of service imports of German multinationals
- Breinlich and Criscuolo

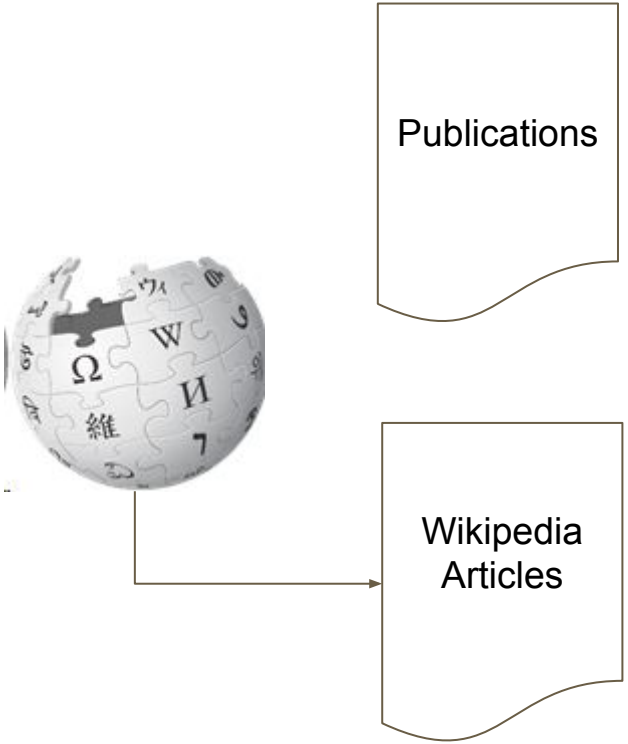
Research Fields

Research Fields

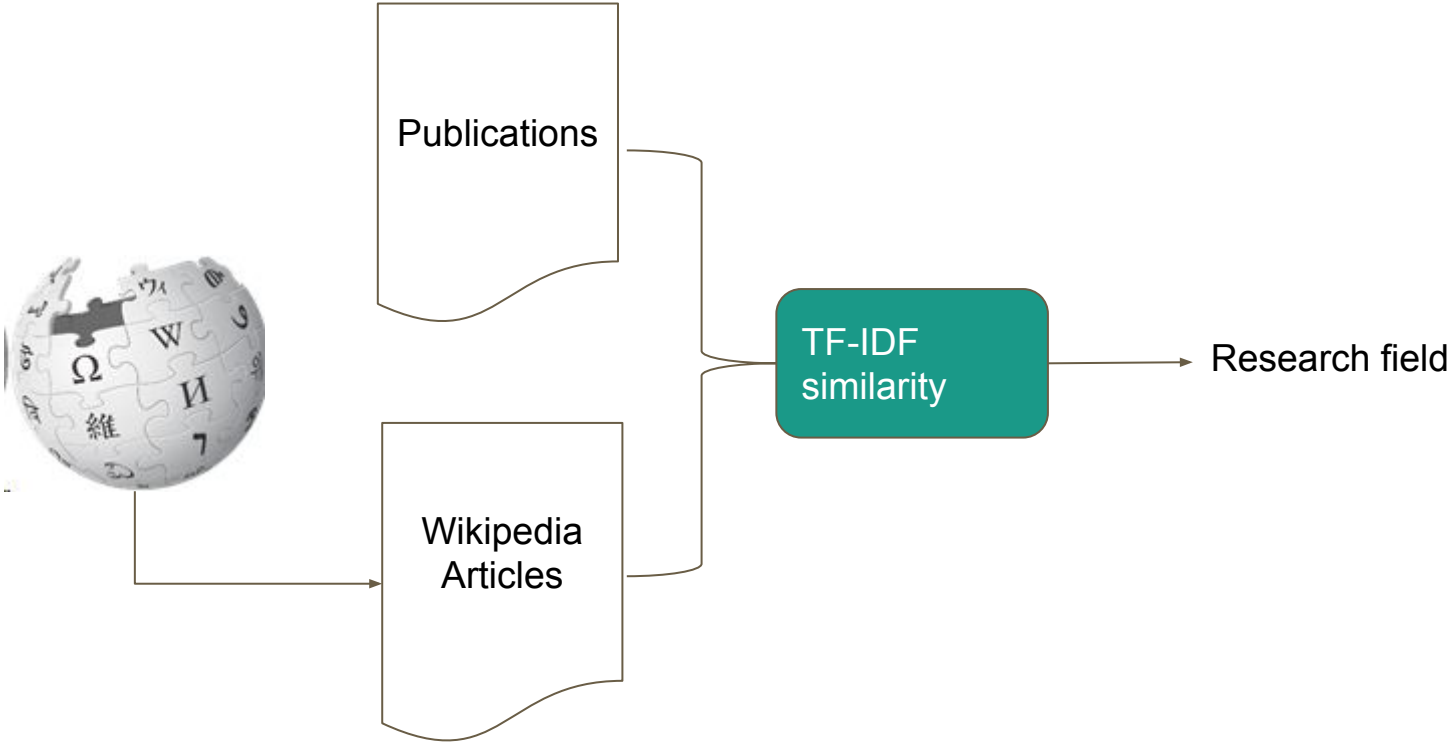
- List of research fields
- Search in Wikipedia the leaves of the hierarchy



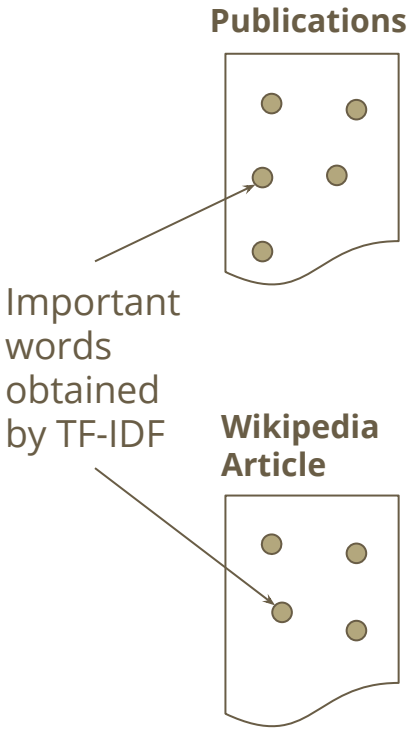
Research Fields



Research Fields



Research Fields



We can **compare** them and discover which **Wikipedia articles** are **similar** to **publications**

Research Fields - Analysis

Document ID	Real R. field	Predicted R. field	Score	Result
893	Medicaid home nursing financing	Health & Social Care, Nursing, Home health nursing management	0.18	✓
2238	Medicine: Body mass index, exercise and inflammatory markers	Health & Social Care, Radiological & Imaging Technologies, Cardiovascular technology	0.107	✗
426	Cognitive rehabilitation of dementia patients	Education Special & Inclusive, Education, Learning Disabilities	0.2	✗

Research Fields - Analysis

- Why does it work?
  - **TF-IDF** is able to select the important words of each article and publication
  - A lot of **research fields** have an **article** in Wikipedia

Research Fields - Analysis

- Why doesn't it work?
  - **Lack** of **articles** about some topics. Eg: "Data-Driven Decision Making in Education" (Edu-5-4)
  - Some topics **share subtopics** so they are similar for TF-IDF

Research Methods

Research Methods

- Our approach to retrieve research methods:
  - NER model
- What is NER?

Apple CEO **Tim Cook** introduces new iPhones at **Cupertino** Flint Center event.

Organization **Person** **Location**

Research Methods

- Why NER?
  - **Research methods** are usually **specific names**, which could be treated as named entities.

    Eg: *snowball sampling, clinical trials, ...*

- Model: Tagger
  - Lample et al., 2016
  - Bi-LSTM-CRF

Research Methods

- Why Tagger?
  - Context influences the meaning of phrases

Eg: *The key advantages of using **content analysis** to analyse social phenomena …*

⇒ ***content analysis*** is a research method.

*Computers are increasingly used in **content analysis** to …*

⇒ ***content analysis*** is **NOT** a research method!!!

Research Methods - Analysis

- 20 random publications analyzed:
  - 12 publications contain at least one right answer
  - However, there is a lot of noise
- Parts of a research method name can appear disjointly
  - Eg: ***Data*** *were* ***collected*** *on both crop-raiding incidents* ⇒ **Data collection**
- Only around **600 research methods** provided. It is **difficult** to **find new research methods** using supervised learning. Other approaches like **semi-supervised** or **unsupervised** learning are **needed**

Challenges

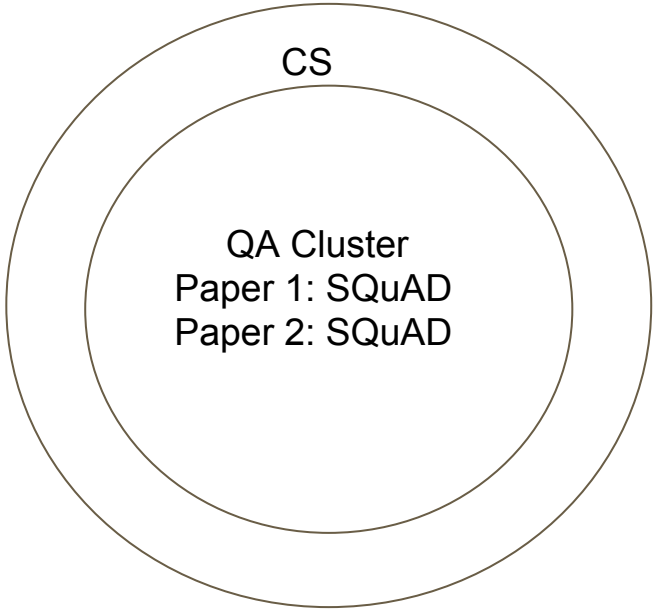
- Not enough training data to **train** an **RC model**
- Difficult to find a **good queries** for dataset retrieval
- Because of this, the result of the **RC model** is **noisy**
- How to identify research methods? What is a research method?

Future Work

- Hypothesis: datasets depends on research fields and vice versa
  - Eg: In the **Question Answering** field (subfield of NLP, CS) the most commonly used dataset is **SQuAD**
  - Eg: 2 papers using SQuAD are likely to be in the same field (QA)

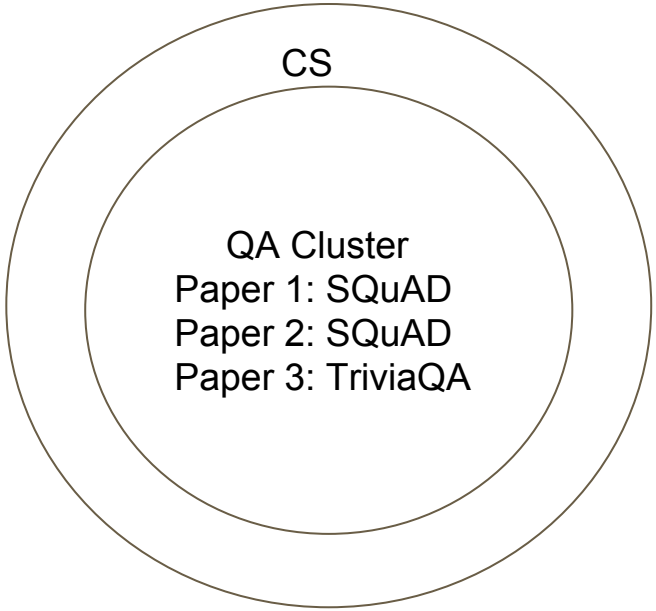
Future Work

- Hypothesis: datasets depends on research fields and vice versa
  - Eg: In the **Question Answering** field (subfield of NLP, CS) the most commonly used dataset is **SQuAD**
  - Eg: 2 papers using SQuAD are likely to be in the same field (QA)
- Build hierarchical **clusters** of papers with the same research **field**



Future Work

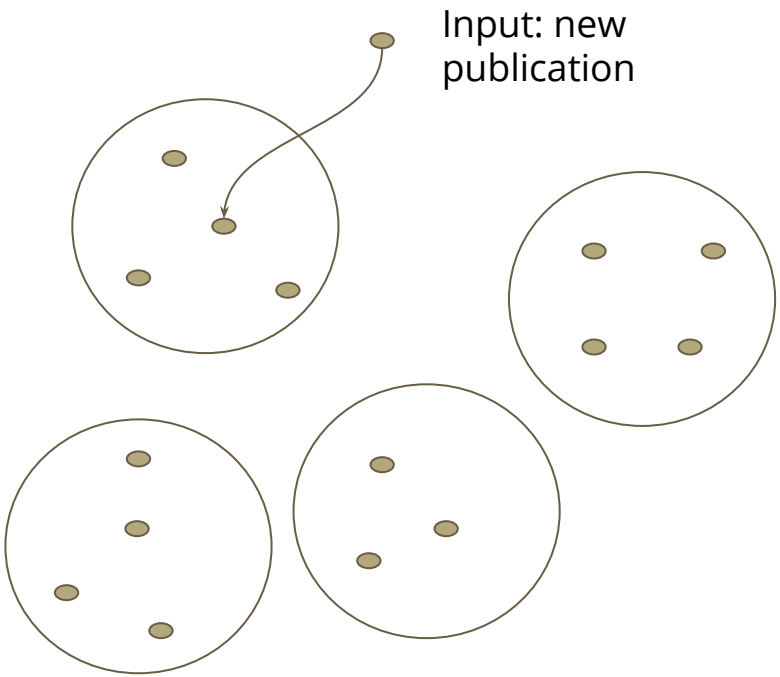
- A **cluster** will have papers with the same research field and **similar datasets**
  - QA cluster will have papers about QA and those papers will use similar datasets like SQuAD and TriviaQA



Future Work

- Recommend to data users similar datasets

- Recommend to data producers fields with small datasets or not enough datasets



Conclusion

- Query Generation Module for dataset retrieval for RC models
- Dataset classifier using entity types and RC scores
- Research field retrieval model using TF-IDF
- NER for research methods retrieval

Thank you