Term Project:
Rich Context Competition

Team 2
20183405 Minho Lee
20174358 Sooji Yoon
20184209 Kuntae Kim
20183309 Giyeon Shin

Contents

Introduction to Rich Context Competition(RCC)

> *"**The goal of this competition is to automate the discovery of research datasets and the associated research methods and fields in social science research publications**. Participants should use any combination of machine learning and data analysis methods to identify the datasets used in a corpus of social science publications and infer both the scientific methods and fields used in the analysis and the research fields."*

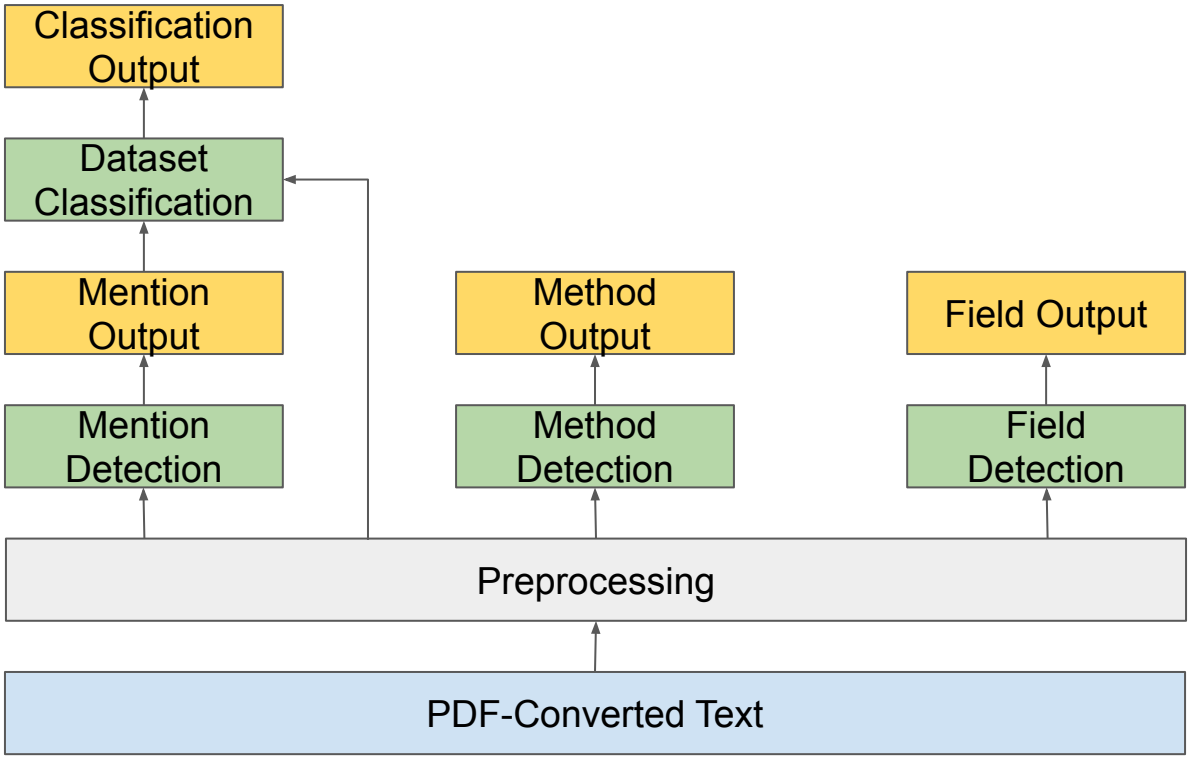
We have to do:

- Task1) Find mentions that may indicate specific dataset
- Task2) For each mention, classify what dataset it is associated to
- Task3) Find research method that the paper used
- Task4) Know what study field this paper is related to

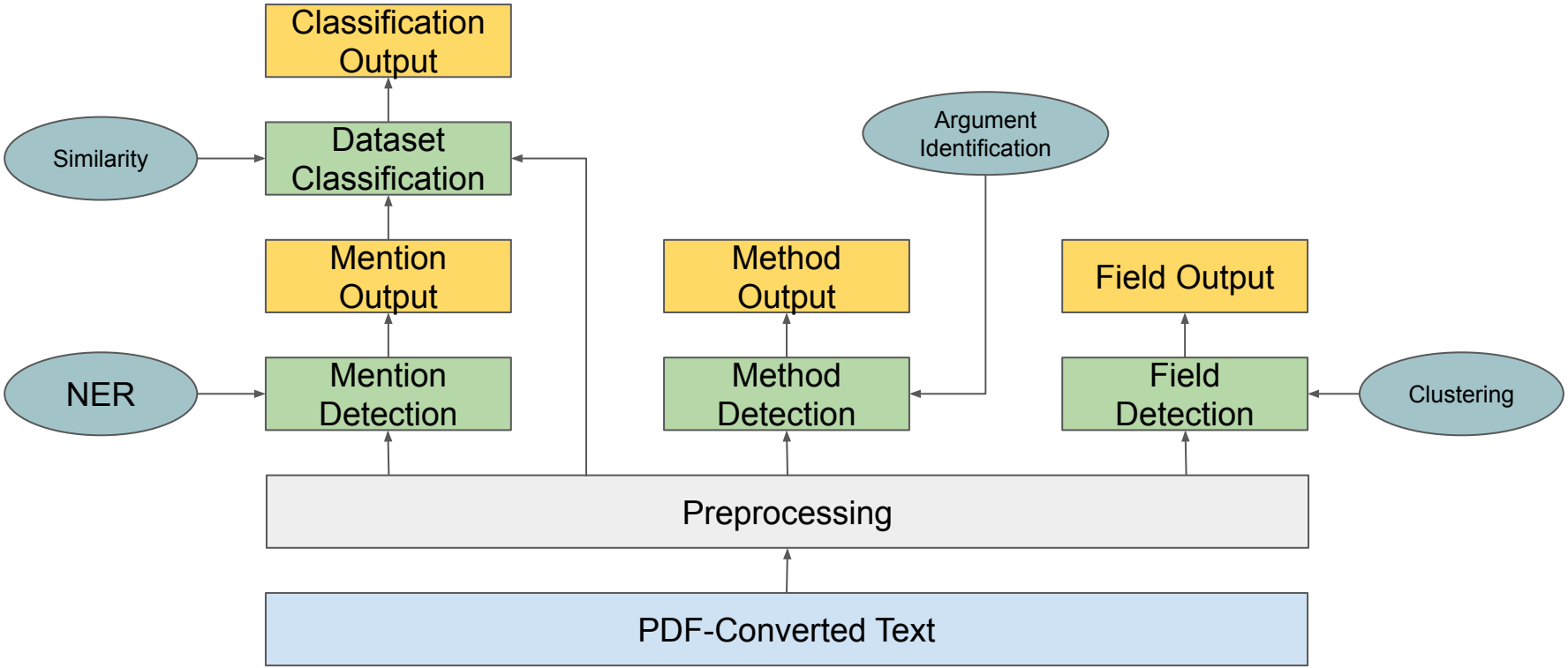
Introduction to Rich Context Competition(RCC)

- Dataset
  - 5000 PDFs and texts of paper for training
  - JSONs which contains metadata of paper - Mentions, Datasets for training
  - Data set indicator
  - Method vocabulary
    - e.g.) t-test, permutation, F-ratio, video interview, ...
  - Research Field vocabulary
    - e.g.) Ecology & Conservation, Physical Geography, Environmental Policy & Law, ...

Our Approach

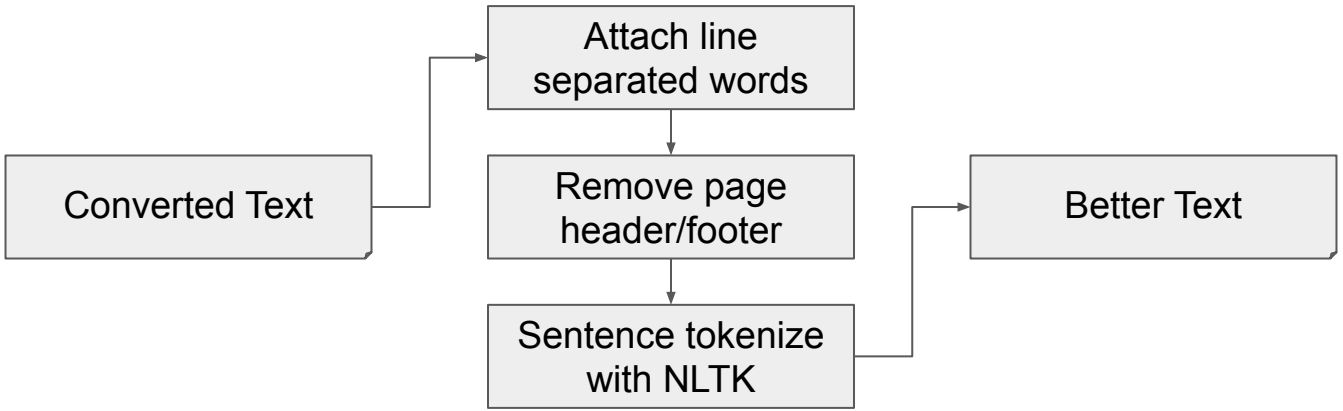


Our Approach

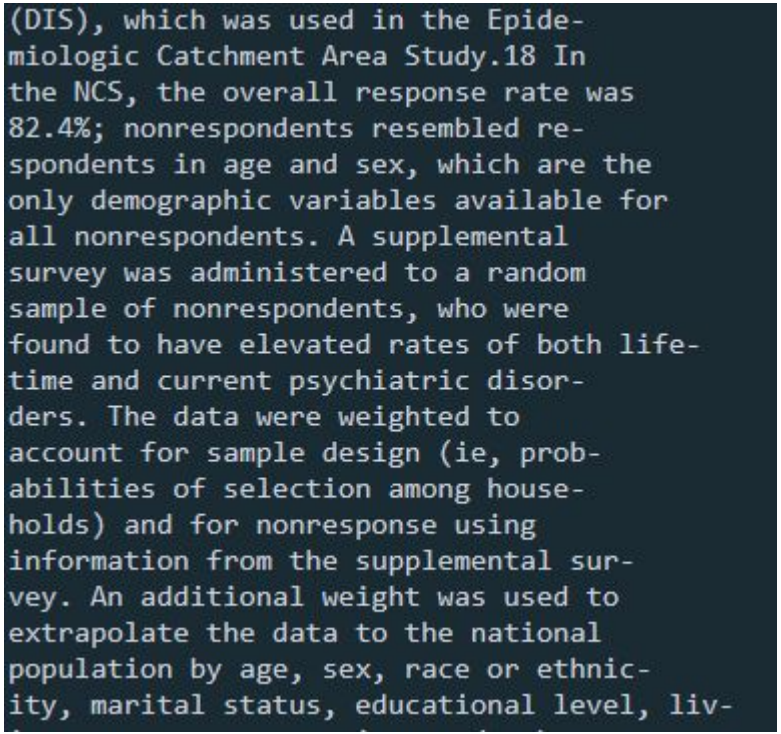


Data Preprocessing

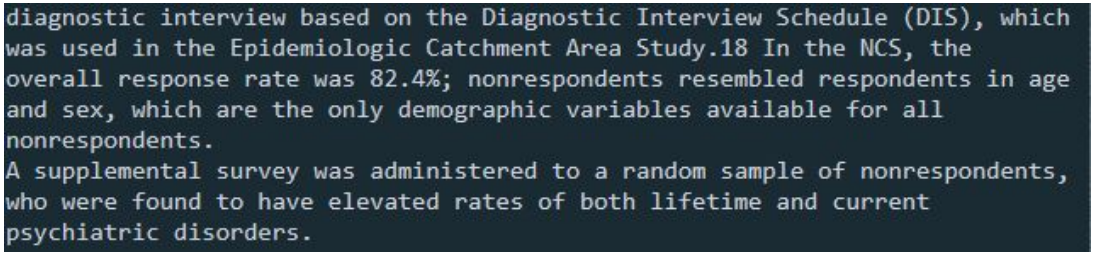
- PDF Converted text is usually ill formed, so try to make it clear



Line-separated text



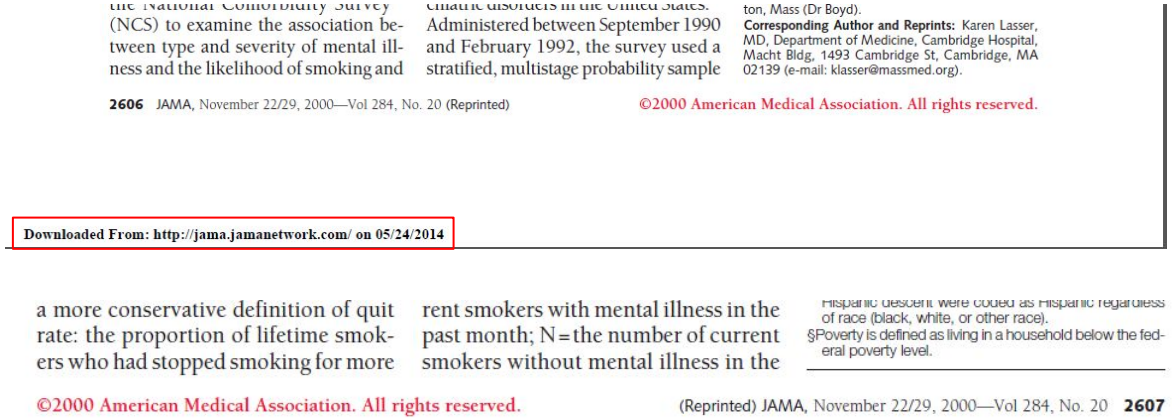
Before



After

Data Preprocessing

- Using PyPDF, get page boundaries
- For each page text, remove common character sequence from front and end



the National Comorbidity Survey (NCS) to examine the association between type and severity of mental illness and the likelihood of smoking and

chiatric disorders in the United States. Administered between September 1990 and February 1992, the survey used a stratified, multistage probability sample

ton, Mass (Dr Boyd).
Corresponding Author and Reprints: Karen Lasser, MD, Department of Medicine, Cambridge Hospital, Macht Bldg, 1493 Cambridge St, Cambridge, MA 02139 (e-mail: klasser@massmed.org).

a more conservative definition of quit rate: the proportion of lifetime smokers who had stopped smoking for more

rent smokers with mental illness in the past month; N = the number of current smokers without mental illness in the

Hispanic descent were coded as Hispanic regardless of race (black, white, or other race).
§Poverty is defined as living in a household below the federal poverty level.

Task1) Dataset Mention Detection Module

Goal & Approach

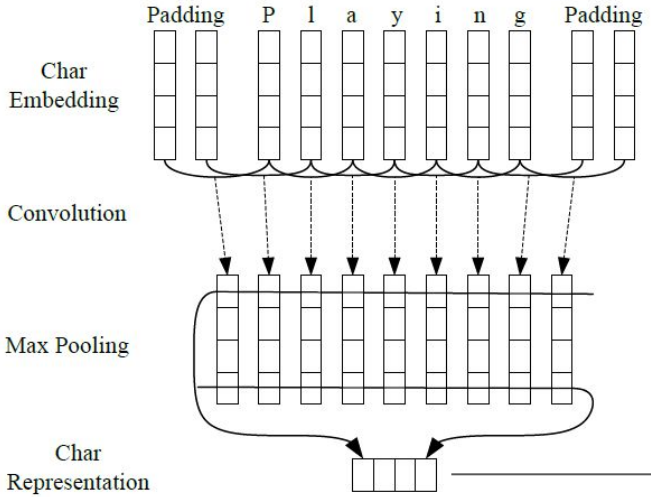
- **Goal: Find mentions that may indicate dataset**
- Approach
  - Used Named Entity Recognition(NER) technique
  - Only aims to find Dataset Mentions

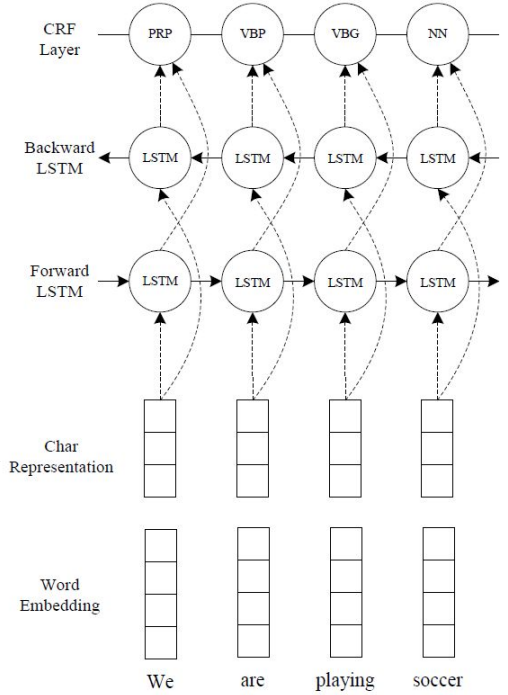
Model

- Deep learning model using CNN, BiLSTM, CRF [1]
- Showing State-of-the-art performance on NER and POS Tagging
- Standard DL Model of sequence labeling and easy to modify

[1] Ma and Hovy, End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, ACL, 2016

Model





Apply CNN at character level to get morphological feature

Apply BiLSTM and CRF to get NER tags

Result

We analyzed data from the Early Childhood Longitudinal Study-Kindergarten Class (ECLS-K).[1] (For a description of the ECLS-K, see http://nces.ed.gov/ecls/pdf/essaysmisc/


```
{
    "publication_id": 1067,
    "mention": "the Early Childhood Longitudinal Study",
    "score": 1
},
{
    "publication_id": 1067,
    "mention": "ECLS-K",
    "score": 1
},
```

Result

- Only trained on mention-positive sentences
- Training:Dev = 9:1
- Dev set score
  - Precision 81.4 / Recall 77.3
  - F1-score 79.3
- Test score
  - We haven’t got yet

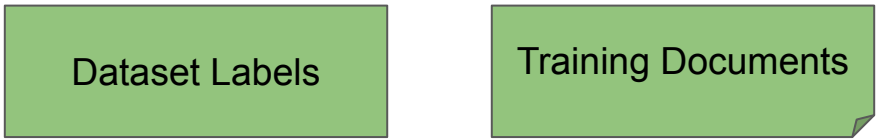
Task2) Dataset Classification Module

Goal & Approach

- **Goal: For each mention, classify what dataset it is associated to**
- Approach
  - Similarity based selection - Doc2Vec, Word embedding(GloVe)

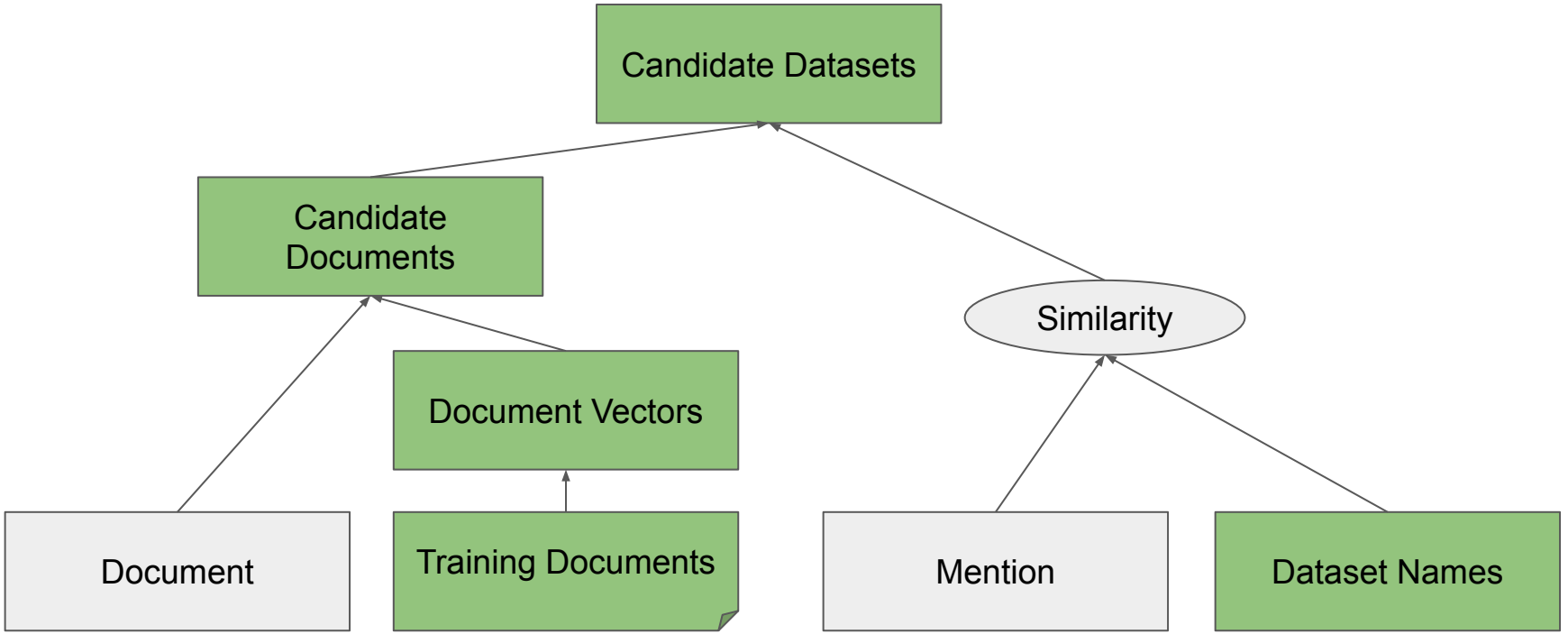
Model

Extracted from Mention Detection Module

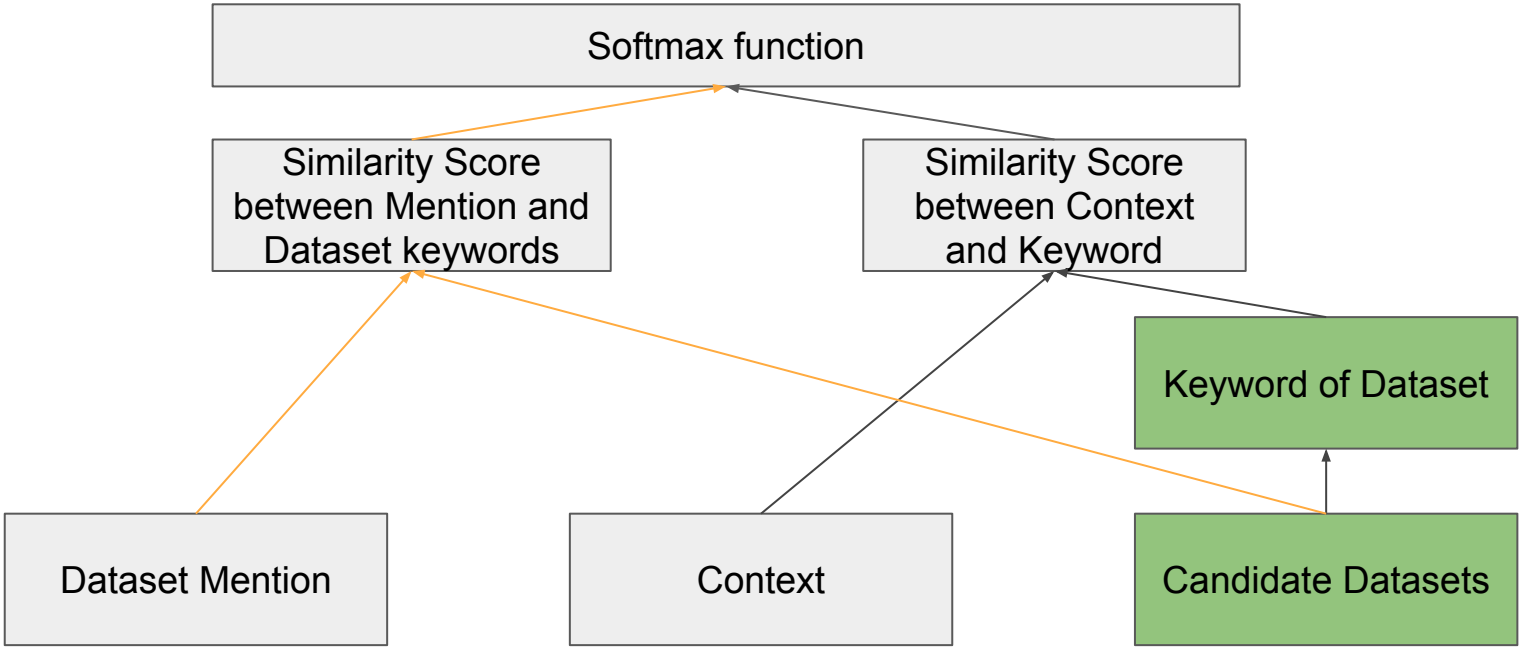


Given Dataset

Model


Candidate Datasets
Candidate Documents
Similarity
Document Vectors
Document
Training Documents
Mention
Dataset Names

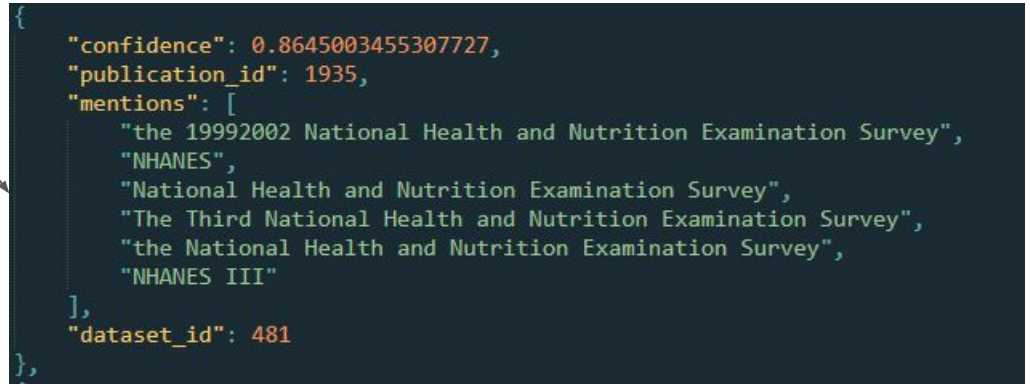
Model


Softmax function
Similarity Score between Mention and Dataset keywords
Similarity Score between Context and Keyword
Keyword of Dataset
Dataset Mention
Context
Candidate Datasets

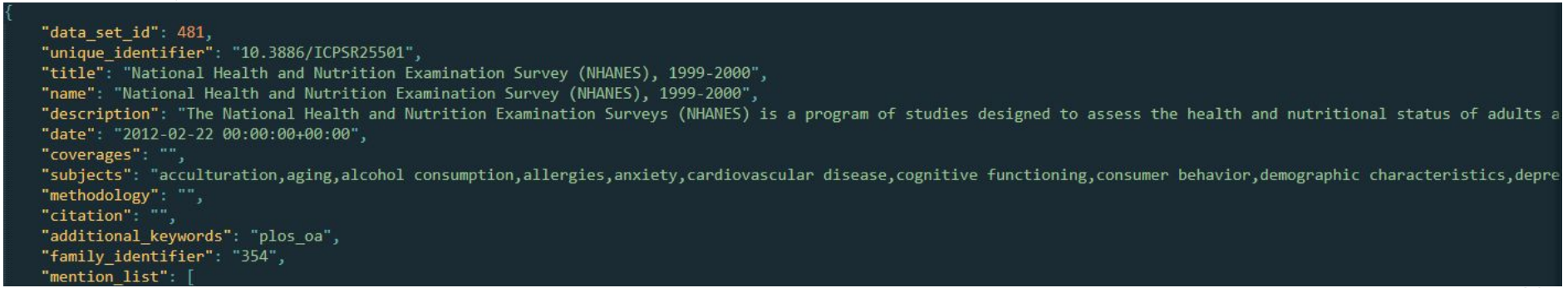
Result

Prediction Example

Dataset Description



```
{
    "confidence": 0.8645003455307727,
    "publication_id": 1935,
    "mentions": [
        "the 19992002 National Health and Nutrition Examination Survey",
        "NHANES",
        "National Health and Nutrition Examination Survey",
        "The Third National Health and Nutrition Examination Survey",
        "the National Health and Nutrition Examination Survey",
        "NHANES III"
    ],
    "dataset_id": 481
},
```



```
{
    "data_set_id": 481,
    "unique_identifier": "10.3886/ICPSR25501",
    "title": "National Health and Nutrition Examination Survey (NHANES), 1999-2000",
    "name": "National Health and Nutrition Examination Survey (NHANES), 1999-2000",
    "description": "The National Health and Nutrition Examination Surveys (NHANES) is a program of studies designed to assess the health and nutritional status of adults a
    "date": "2012-02-22 00:00:00+00:00",
    "coverages": "",
    "subjects": "acculturation,aging,alcohol consumption,allergies,anxiety,cardiovascular disease,cognitive functioning,consumer behavior,demographic characteristics,depre
    "methodology": "",
    "citation": "",
    "additional_keywords": "plos_oa",
    "family_identifier": "354",
    "mention_list": [
```

Task3) Method Finding Module

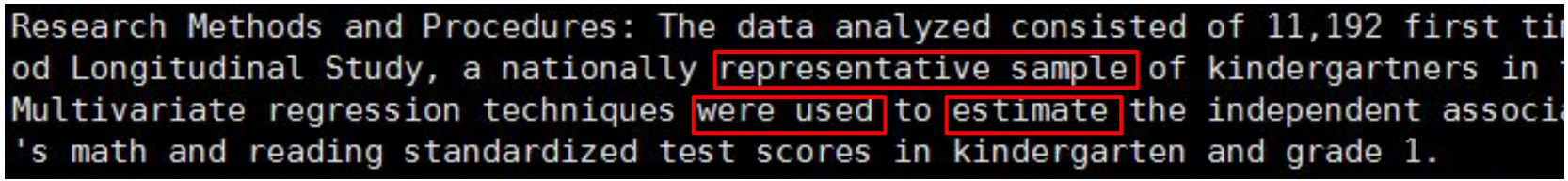
Goal & Approach

- **Goal: Finding research methods that the paper used**
- Approach
  - In the paper, some sentences involving specific verbs describes research methods.
  - Using Semantic Role Labeling (SRL), we find out those specific predicates and finally take research methods.
- Model
  - End-to-End technique to predict jointly the predicates and arguments in Argument Identification Task.[1]

[1] Luheng He, Kenton Lee et al., Jointly Predicting Predicates and Arguments in Neural Semantic Role Labeling, ACL, 2018.

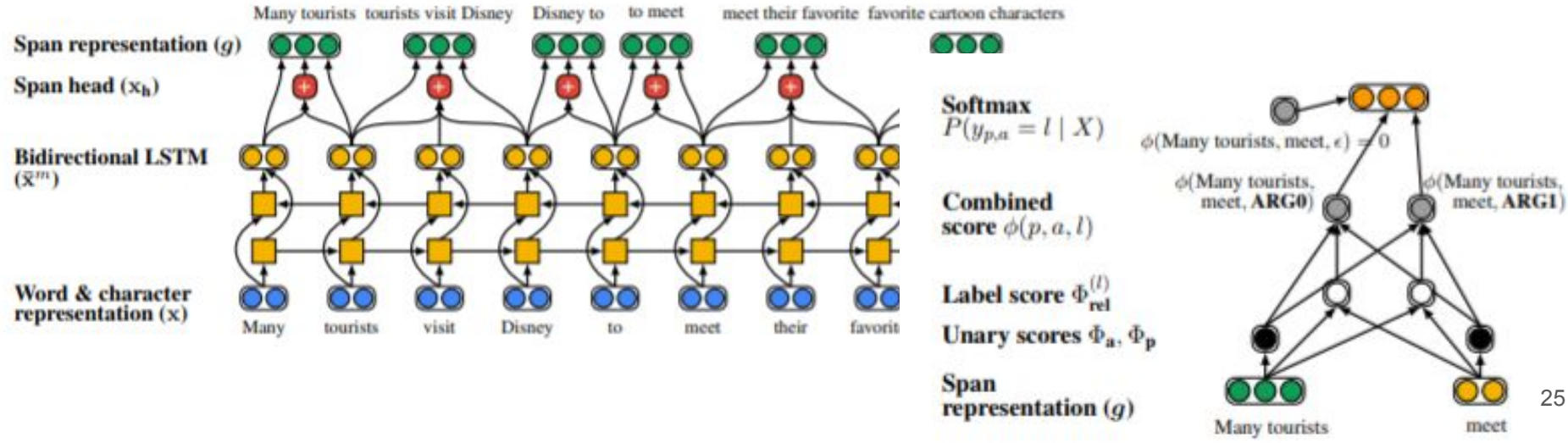
Approach

- Why do we Choose Semantic Role Labeling?
  - This task aims to find methods not only given in dataset but also unseen methods.
  - Some specific verbs means methods used in the publication (e.g use, provide).
  - Therefore, a methodology for jointly predicting verbs and methods is required.



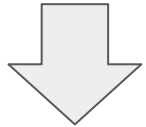
Model

- How to apply this model in this task?
  - Constructing silver input datasets using normalized text and given metadata.
  - Labeling methods appeared in the metadata and verbs in the sentence which methods appear
  - Learn the model through the labeled silver input dataset



Result

```
Article Introduction Structural symbolic interactionism emphasizes the impact
n with others to convey who they are, or the meanings of their identities.
Self--consisting of multiple identities--emerges from the patterned and organ
& Stets, 2009).
In contrast, traditional symbolic interactionism opposes any suggestion that
e fluid as individuals construct identities differently across situations.
Along the lines of structural symbolic interactionism, identity theory specif
```



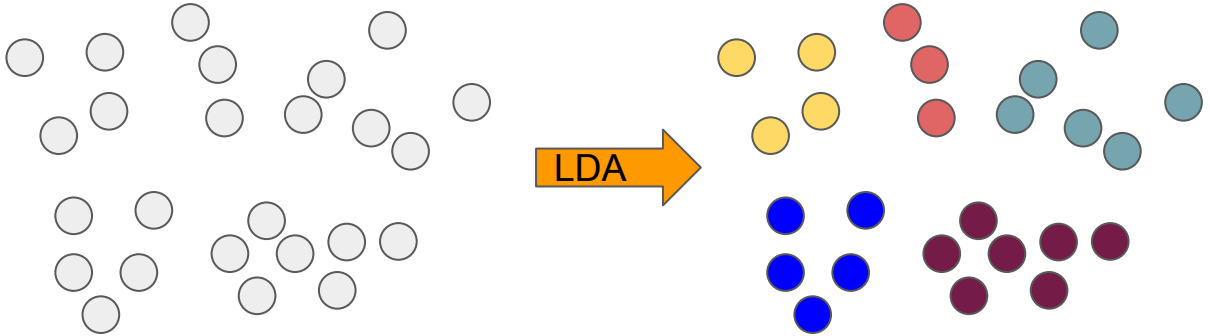

```
{
        "method": "access",
        "publication_id": "5134",
        "score": 1.0
},
{
        "method": "symbolic interactionism",
        "publication_id": "5134",
        "score": 0.8
},
```

Task4) Field Finding Module

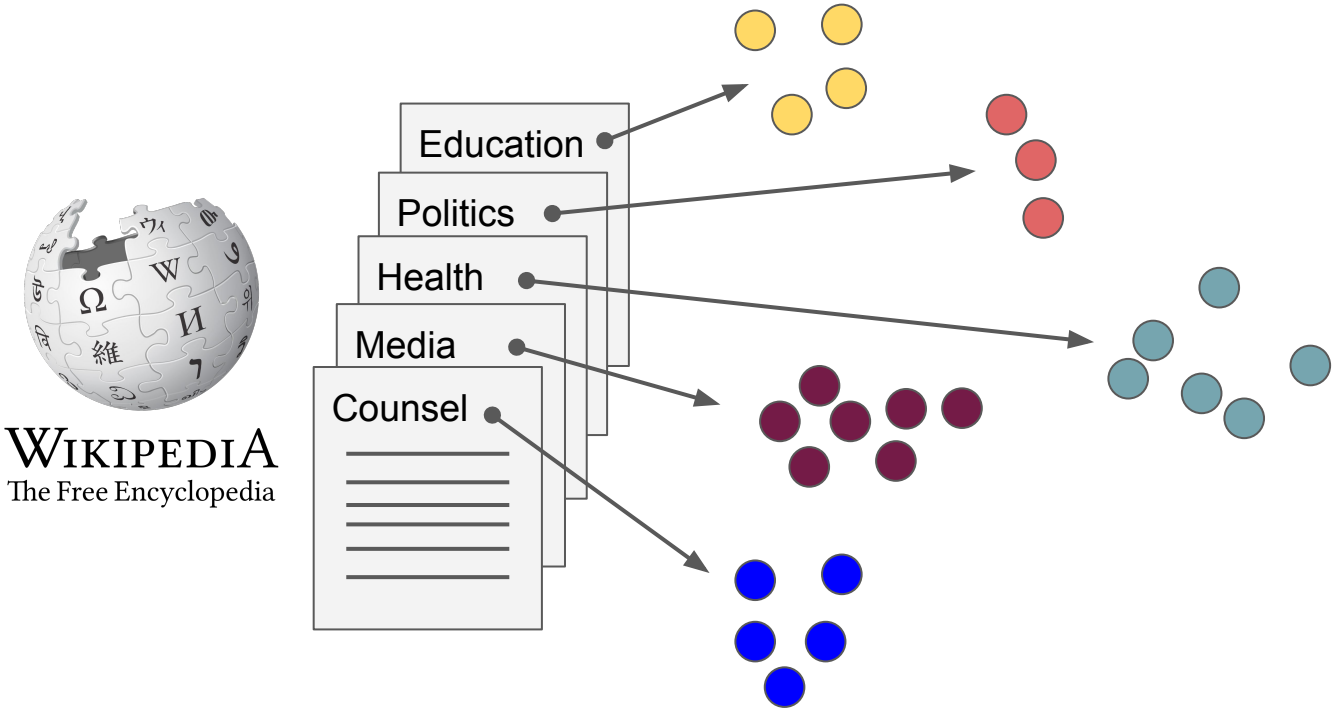
Goal

- **Know What study field each paper is related to**
- Challenges
  - The given documents are not annotated
  - We have only a list of research-fields i.e., no information about Social Science domain research fields
- Solutions
  - Clustering documents with auto calculated topics
  - Searched research-fields in the list on Wikipedia, and get the documents

Overview - Clustering



Overview - Mapping



Clustering-LDA

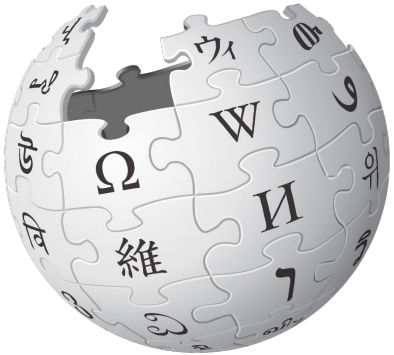
- Preprocessing
  - Remove stopwords,numbers, and symbols, Set minimum frequency=5
- We used first 50 lines in each document
  - The authors usually describe the research field in the section of abstract or introduction

Example of Clustering Result

Topic1	The epidemiology of **panic attacks**, **panic disorder**, and **agoraphobia** in the **NCS-R**
	Lifetime prevalence and age-of-onset distributions of **DSM-IV disorders** in the **NCS-R**
	Failure and delay in initial treatment contact after first onset of **mental disorders** in the **NCS-R**
	Lifetime and 12-month prevalence of **bipolar spectrum disorder** in the **NCS-R**
Topic2	**Medicaid Cost** Containment and Access to **Prescription Drugs**
	Public support for policies that would help people with **chronic conditions**
	The effects of health changes on projections of **health service** needs for the elderly population of the United States
	Data Watch: Victim costs of violent crime and resulting **injuries**
Topic3	Unmet Need for **Personal Assistance Services**: Estimating the Shortfall in Hours of Help and Adverse Consequences
	Volunteerism and **Socioemotional Selectivity** in Later Life
	Incidence of **Four-Generation Family Lineages**: Is Timing of **Fertility** or **Mortality** a Better Explanation?
	Church-based **social support** and **mortality**

Mapping

- For mapping each clusters to a study fields, we compare document similarity among papers in a cluster and documents which explains fields.
- How to get documents explaining study fields?
- How to compare each documents?



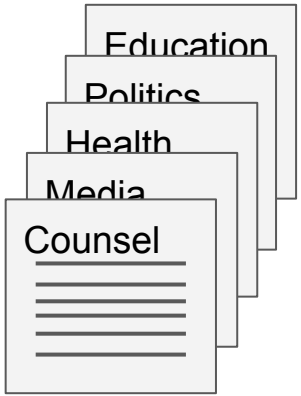
WIKIPEDIA
The Free Encyclopedia

Doc2vec paragraph embeddings

Mapping - Doc2Vec

- Search study fields in wikipedia and define found pages as documents of field.
- Train doc2vec model about above field documents.
- For each clusters,
  - Infer papers' doc2vec embedding vectors.
  - Rank Top 10 most similar fields for each papers by comparing document similarities.
  - Assign a cluster's field as the most Top 10 ranked field.

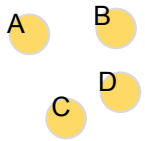
Mapping - Doc2Vec



PaperA: 1. Counsel 2. Media 3. Health ...
PaperB: 1. Health 2. Media 3. Politics ...
PaperC: 1. Media 2. Health 3. Education …
PaperD: 1. Health 2. Education 3. Counsel

Compare Similarities

Cluster ID: 1



Health	4
Media	3
Counsel	2
Education	2
Politics	1

Example of Result

Maternal & Child Health

1164 / Maternal depressive symptoms, father's involvement, and the trajectories of child problem behaviors in a US National Sample
1175 / The ecology of childhood overweight: A 12-Year longitudinal analysis
1670 / Shortened sleep duration does not predict obesity in adolescents
1909 / Young adult outcomes of children growing up with chronic illness: An analysis of the National Longitudinal Study of Adolescent Health

Social Psychology

2350 / The role of drinking in new and existing friendships across high school settings
3344 / Signs of Social Class: The Experience of Economic Inequality in Everyday Life
351 / Perceived interpersonal mistreatment among obese Americans: Do race, class, and gender matter?
3665 / Shyness, Self-Construal, Extraversion–Introversion, Neuroticism, and Psychoticism
3717 / Young Women With Anorexia Nervosa
3813 / Children’s Attitudes and Stereotype Content Toward Thin, Average-Weight, and Overweight Peers

Summary

- For each task, make individual module
- Task1) Dataset Mention Detection
  - Use Named Entity Recognition technique
- Task2) Dataset Mention Classification
  - Candidate finding by calculating similarity
- Task3) Study Method Recognition
  - Predicates and Methods finding by Arguments Identification technique
- Task4) Study Field Recognition
  - Clustering documents - LDA
  - Mapping clusters and research fields by calculating distances

Reference

- NER: Ma and Hovy, End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, ACL, 2016
- Argument Prediction: Luheng He, Kenton Lee et al., Jointly Predicting Predicates and Arguments in Neural Semantic Role Labeling, ACL, 2018.