Coleridge Initiative - Show US the Data

Simple and Strong Baseline

Mikhail Arkhipov

Third Place Solution

Final Leaderboard

ZALO FTW	0.576
Chun Ming Lee	0.575
Mikhail Arkhipov	0.558
OsciiArt resistance0108	0.513
Kramarenko Vladislav	0.486

The Task

Given a collection of texts from scientific publications extract all mentions of datasets from each text. Mention is a substring from the raw text.

Example:
The newly released Early Childhood Longitudinal Study offers the data needed to examine these processes within a national cohort of students beginning in elementary school.

Training Data

The definition of the dataset is given by 45 distinct positive examples which mentions was labeled in 14316 documents. All of them are capitalized (words starts with a capital letter). Each document contains 350 sentences on average.

Simple Capitalization Pattern Matching

All sequences in the training labels consist of capitalized words (words that starts with capital letter). A straightforward approach could be extracting such sequences from the given texts.

Matching and missing examples from the train labels:

Matches

- National Education Longitudinal Study
- World Ocean Database
- High School Longitudinal Study

Missing

- Baccalaureate and Beyond
- COVID-19 Image Data Collection
- Aging Integrated Database (AGID)
- Program for the International Assessment of Adult Competencies

Fixing missing cases

Add the following items as possible connection words:

- Prepositions
- Conjunctions
- Brackets
- Hyphens
- Numbers

This covers pretty much all training label cases.

False Positives

False Positive Matches

- Supplemental Materials
- Rhoades & Torres-Olave
- El Niño and La Niña
- United States and the United Kingdom
- Type I and Type II
- American Indians and Alaska Natives

True Positive Matches

- COVID-19 Image Data Collection
- National Education Longitudinal Study
- Aging Integrated Database
- School Survey on Crime and Safety
- World Ocean Database
- COVID-19 Open Research Dataset

Keywords

False Positive Matches

- Supplemental Materials
- Rhoades & Torres-Olave
- El Niño and La Niña
- United States and the United Kingdom
- Type I and Type II
- American Indians and Alaska Natives

True Positive Matches

- COVID-19 Image Data Collection
- National Education Longitudinal Study
- Aging Integrated Database
- School Survey on Crime and Safety
- World Ocean Database

COVID-19 Open Research Dataset

In many cases, labels from the train set include words that seems to be a strong indicator of a dataset mention.

Keywords

I come up with the following list of keywords:

- Study
- Survey
- Assessment
- Initative
- Data
- Dataset
- Database

Non-Dataset matches

Still there are a lot of mentions that seems to be referring to non-dataset entities:

- National Assessment Governing Board
- Agency for Healthcare Research and Quality Data Center
- Office of Coast Survey Field Procedures Manual
- Program Assessment Rating Tool
- Florida Comprehensive Assessment Test

Stopwords

Manual analysis of the results resulted in the following list of stopwords:

Administration	Agency	Association	Board	Bureau
Center	Centre	Clearinghouse	Committee	Company
Consortium	Corps	Department	Division	Documentation
Format	Foundation	Framework	Group	Institute
Lab	Manual	Office	Organisation	Organization
Rating	Scale	Service	Supplement	System
Test	Tool	University	Variables	

If any of these stopwords present in the mention, the mention is discarded.

Brackets with Abbreviations

Even with filtration by key/stop words there are a lot of false positives like:

- Case Study 4)
- A Comparative Study From
- Programs of Study
- $\frac{1}{4}$ 57) This Study Consensus

To mitigate such cases the pattern is extended to end with brackets containing abbreviations. For instance:

- Labour Force Survey (LFS)
- New York City Community Air Survey (NYCCAS)

Co-occurrence Statistics

A possibly good indicator of a dataset is the frequency of co-occurrence with the word "data". The relative frequency of a substring *str* is defined as

$$F_d = \frac{N_{data}(str)}{N_{total}(str)}$$

where $N_{data}(str)$ is the number of times the *str* occures with "data" word (parentheses are dropped) and $N_{total}(str)$ is the total number of times *str* present in texts.
All mentions with $F_d < 0.1$ are dropped.

Prediction Phase

Using all filtering methods mentioned the collection of detected mentions is used as dictionary for substring search during prediction phase. No postprocessing used (e.g. removing punctuation, lowercasing). Brackets are dropped.

Additionally, for each detection, we look for parentheses after the detected mention. If there are parentheses and some abbreviations inside, these abbreviations are added as a separate predictions.

Final Solution Summary

- Find all capitalized sequences (mentions) followed by parentheses with abbreviations inside for both train and test documents
- Remove mentions without keywords
- Remove mentions with stopwords
- Remove mentions that occur with the word "data" less than 10% of the time
- Find all occurrences of the detected mentions in test texts

Other Attempts

- Hearst Patterns motivated templates
  - "data/samples/obtained from Xxx Xxx"
  - "datasets: Xxx Xxx, Xxx ..."
  - "Xxx Xxx and Xxx Xxx datasets"
- Children of the "dataset" word from the dependency tree
- Titles of the papers from arXiv of the form "Xxx Xxx ...: dataset ..."

Motivation for Simple Pattern Matching Solution

In the low-data regime there are some evidence that simple pattern matching works well.

In Entity Linking (extract substring and point to the database page), dictionaries extracted from Wikipedia hyperlinks is a very strong baseline[1].

In Hypernimy Detection, Hearst Patterns like "xxx like yyy and zzz" are still hot[2].

[1] Joint Multilingual Supervision for Cross-lingual Entity Linking
https://arxiv.org/abs/1809.07657

[2] Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora
https://arxiv.org/abs/1806.03191

SPASIBO