Coleridge Initiative - Show US the Data
2nd place solution

Chun Ming Lee

Agenda for today

- My background
- Summary of my approach
- Training approach
- What didn’t work
- Important findings

Chun Ming Lee

- NLP Data Scientist at various Singaporean startups
- Management Consultant at McKinsey & Co.
- Software Developer at Bermudan Hedge Fund
- MBA, London Business School
- BSc. Computer Science, Carnegie Mellon University
- Kaggler
  - 1st place winner, 2020 Jigsaw Multilingual Toxic Comments NLP competition
  - 1st place winner, 2018 Jigsaw Toxic Comments NLP competition

Defining the task

Task: Identify datasets within scientific publications

Uncertainty about definition of dataset e.g., Intergovernmental Panel on Climate Change (IPCC)

Metric: Jaccard-based FBeta score (with beta = 0.5)

False positives penalized more than False negatives

Data provided: ~14,000 papers with partial labels

Overlap of papers with public test-set

Summary of my approach

Key insight: Leverage formal academic writing style to identify datasets

Excerpt from sample publication

Using a serum T level o325 ng/dl, the **Baltimore Longitudinal Study of Aging (BLSA)** reported that approximately 12, 20, 30, and 50% of men in their 50s, 60s, 70s, and 80s, respectively, are hypogonadal.
…
In the **BLSA**, the average decline was 3.2 ng/dl per year among men age 53 years at entry.
…
A number of longitudinal epidemiologic studies, including the **Baltimore Longitudinal Study of Aging**, the New Mexico Aging Process Study, and the Massachusetts Male Aging Study, have demonstrated age-related increases in the likelihood of developing hypogonadism.
…
Signs of hypogonadism can be effectively treated using **testosterone replacement therapy (TRT)**.

Datasets frequently mentioned by *LONG-FORM (ACRONYM)*

Referred to only by acronym

Also referred to without acronym

Not all *LONG-FORM (ACRONYM) strings are datasets*

Model flow

1. Search

Search for strings in the format *LONG-FORM (ACRONYM)* using the Schwartz-Hearst text search algorithm e.g., *Baltimore Longitudinal Study of Aging (BLSA)*

2. Classify

Filter long-form part of strings from **(1)** using fine-tuned roberta-base Transformer
- **Accepted:** *Baltimore Longitudinal Study of Aging*
- **Rejected:** *testosterone replacement therapy*

3. Threshold & Propagate

For accepted candidates from **(2)**
- Search for long-form part of candidate (*Baltimore Longitudinal Study of Aging*) across documents
- Accept acronym if separately mentioned in same document (*BLSA*)

Training approach

Train roberta-base Transformer with binary classifier head using ~5K manual annotations

Excerpt from annotations

```
Global Ocean Data Assimilation System,0
Common Data Model,0
Graduation Rate Survey,1
International Surface Pressure Databank,1
Academic Libraries Survey,1
Global Data Assimilation System,0
Current Population Surveys,1
Breeding Bird Survey,1
Image and Data Archive,1
Coast Survey Development Laboratory,0
Soil and Water Assessment Tool,0
National Pupil Database,1
National Snow and Ice Data Center,0
Teacher Followup Survey,1
Federal Geographic Data Committee,0
Electronic Data Interchange,0
Computer Assisted Data Entry,0
Curriculum and Assessment Policy Statement,0
Continuing Survey of Food Intakes by Individuals,1
Sequence Read Archive,1
National Air Toxics Assessment,1
```

Label as reject for ambiguous candidates given heavier penalty on False Positives

What didn't work

Attempts to use semantic context worsened performance

Designed for short texts, hard to adapt to publication-length texts

QA models

Tested Transformer Question Answering (QA) models, pairing publication text with question "What is the dataset?"

- Frozen QA models performed better than fine-tuning on training data
- Pretrained QA models often flag "obviously" incorrect strings (e.g., Food & Drug Administration) raising questions about what they've learnt

NER models

Tested Transformer Named Entity Recognition (NER) models with datasets annotated as a Named Entity type

- Similar issue with QA models, flagging organization names etc.

Curated data

Selecting subset of sentences in publications for training above QA/NER models

- **Using data from the National Longitudinal Survey of Youth of 1997**, the purpose of this thesis is to investigate whether and how parental divorce affects children's post-baccalaureate educational attainment.

Key takeaways

1. Simple models using domain knowledge often beat complex general models
2. Annotation methodology often more important than modelling
3. Trendy Transformer QA models perform poorly in real-world tasks

Q&A

Thank you for your time!