Data Science at Elsevier

12/26/20 10:10 PM

Speaker: Brad Allen and Daniel Kershaw (Elsevier)
Date: 2020/9/15

A. Company Overview

From content publishing to data solutions
Read -> Search -> Do

Five main customer segments
- Clinicians: "Consider this treatment for this patient"
- Researchers: "This article answers your questions"
- Governments: "This is the research to invest in"
- Pharmaceutical companies: "This is the cancer treatment that you should pursue"
- Nursing students: "This is the area you need to improve to qualify"

Challenges that customers face
- Global research spend is growing, however it might not be the same for efficiency
- Researchers lack the tools they need to be effective
  - 70% - 80% of research asks the wrong questions
- Life-saving drugs are expensive to develop, success rate is very low
- Health providers cannot save lives without the best information
  - Preventable medical errors is the 3rd largest cause of death in the U.S.

Data Science matters
- Augment, not replace, professional decision making in science and medicine
- Enable better outcomes through
  - Delivery of timely, appropriate advice to help perform routine tasks quickly and accurately
  - Enhance discovery and query over massive amounts of information

Data Sources
- Elsevier content and metadata
  - Books, peer reviewed journal articles, medical literature (guidelines)
- Data derived from content
  - Citation data, organizational profiles, personal profiles, funding profiles
- Data derived from usage
  - Article views, access statistics
- Data obtained from third parties
  - Open data / knowledge resources on web, data from healthcare partners

Vision
- A data platform for outcome-improving insights
  - Acquire content -> Extract, link & curate data -> Deliver insights
- A linked data model of science & its social graph
  - Extract attributes and relationship of various entities (trails, treatments, articles, patents, practitioners, funders, etc..)
- Challenge: building bottom-up through data sharing & governance
  - Make sure all effort are in one accord across different assets
  - Asset types include (not limited to): data models, datasets, catalogs, data sources, data processors (transformers, annotators, classifiers), data flows (compositions of data sources, connectors), software libraries (ETL, graph algorithms, NLP libraries)

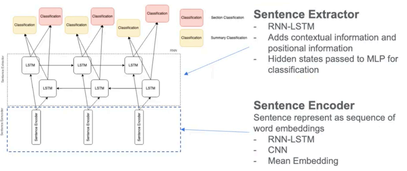
B. BAU (Business as Usual) Example: Summarization at scale

Focus: extraction - select the best sentences which summaries the document

Training data - rouge sampling
Rouge: Recall-Oriented Understudy for Gisting Evaluation

- Author highlights: more concise bullet points about the article
- Sample greedily sentences which are most like the author highlights (in decreasing order of contribution)
  These are positive examples

The model architecture
- A sentence encoder layer first, then sentence extractor layer
- Encoder layer: get a sequence of embeddings
- Extractor layer: binary classification into summary sentence (pos) or not summary sentence (neg), using LSTM



** Why not using a BERT (or transformer in general)?
- Much more interpretable
- Also more maintainable

Training extractive model
- Word embeddings: GloVe 100
- Learning rate 0.0001, drop-out 0.25, trained up to 50 epochs with early stopping

Numeric results
Evaluation metrics: rouge-l-f@4 scores for each of the six variations in sentence encoders

- **Scores** for CNN based (best performance compared to MAN and RNN) sentence encoder
  -

Shuffled	Biology	Computing	Economics	All
False	21.03	22.97	21.61	22.19
True	20.11	22.62	21.09	20.75

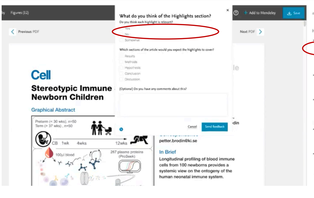
rouge-l-f@4 scores for the CNN based sentence encoder model without additional features.

- **Human evaluation** for the predicted summary sentences (scale 1-4)

Discipline	Dive	Info	Simp	Rele
Bio Science	2.81	2.49	2.93	2.65
Computing	2.95	1.57	2.95	1.86
Economic	3.00	2.67	3.67	2.33
All	2.90	2.37	2.83	2.51
Author highlights	3.01	2.78	2.81	3.2

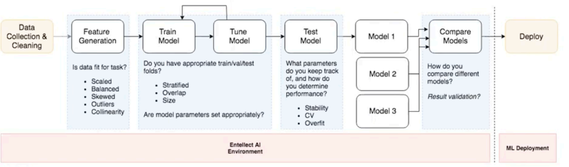
Results for human evaluation

Production
- Summary sentences are listed on the top as "Highlights"
- Also contains an interaction "what do you think of this highlight" to collect feedback
-



C. Next steps

- *Roofshot: continuous modeling and quality control in data solution production workflows*
  -



- *Moonshot: question answering for knowledge graph construction*
  - Using question answering as a general technique for knowledge extraction from free text content
  - Using QA over the scientific literature can lead to the automatic creation of knowledge graphs for science and machine
  -

- *Open source data - 40K OA articles*
  "Elesevier OA CC-By Corpus", paper available
  - JSON format, full text with references, machine readable text
  - 60% contains author highlights
  - Helps to develop new datasets from there