The Economics of Artificial Intelligence and Machine Learning for Semantic Enrichment

Tuesday, 4 December 2018, London
STM Tools and Standards Seminar
10:15 AM Ignite Session 1

Jay Ven Eman, Ph.D., CEO
Access Innovations, Inc. / Data Harmony
j_ven_eman@accessinn.com
www.accessinn.com
+1.505.998.0800
Albuquerque, NM USA

Access Innovations, Inc.
The Science behind the Semantics™
www.accessinn.com
Summary

- AI/ML/DL hold promise for STM
- But big, headline grabbing failures
- Costs can run to the billions
- Choose carefully
- Choose narrowly
- Focus on improving content for customer utility and process workflow improvements
AI – It’ll Be Awesome!
Headlines – the good, the bad, & the ugly

- **The Good**
  - “Can Artificial Intelligence Help Reduce False-positive Mammograms?”
  - “You Might Want Artificial Intelligence Reading Your Next Mammogram”
  - “When AI writes the Court Rulings”
  - “Fast and Accurate Annotations of Short Texts with Wikipedia Pages”
Headlines – the good, the bad, & the ugly

- The Bad
  - “Without Humans, Artificial Intelligence is Still Pretty Stupid”
  - “The Future of AI Depends on a Huge Workforce of Human Teachers”
  - “Google Has Picked an Answer for You – Too Bad It’s Often Wrong”
  - “Artificial Intelligence Still Isn’t a Game Changer?”
  - “Google, Smoogle. Reference Librarians Are Busier Than Ever”
Headlines – the good, the bad, & the ugly

- The Ugly

- “Some AI Lessons from Watson’s Failure at MD Anderson”
- “MD Anderson Benches IBM Watson In Setback for Artificial Intelligence in Medicine”
- “Artificial Intelligence and Bad Data”
- “Sky-high Salaries Are the Weapons in the AI Talent War”
- And for STM – SciGen – “Tech society retracts 29 articles, ousts three editors for ‘systematic violation’ of peer review polices”
“Teaching a machine to read a record is a lot harder than anyone thought.”

First, know that Watson was getting good results!

The Fail
- Not enough data
- Inconsistent and bad data
- Incompatible systems (Watson ↔ EPIC’s EHR)
- Changing objectives → oops, need to retrain Watson!
- Lack of AI knowledge and expertise
- Cost overruns – US$62 million spend before tabling project
Get back to business basics

- “Don’t believe the hype – AI is just a tool at the end of the day, but a very clever tool…”

- The Hype
  - “…find growth and accelerate innovation within an open data environment”
  - “…breaking the silos of the status quo…”
  - “Adopting a holistic data strategy…”
  - “…providing next generation…”
  - “IBM Watson capabilities to unlock previously unavailable data insights”
Get back to business basics

- Stick to basic business practice
  - Use cases
  - Business cases
  - “Plan your dive. Dive your plan.”
  - Without a big budget, keep your expectations in check – narrow your focus
    - Do you have US$62 million to blow?
Get back to business basics

- Use cases
  - Anti-SciGen, fake news detection, submissions analysis
  - Auto text generation and summarizations (big in the news business)
    - Court rulings (e.g. Prometea – Argentina)
    - Washington Post, Associated Press (e.g. sports summaries)
  - Machine automated indexing (MAI), semantic enrichment
  - Image analysis and recognition, info-graphics
  - Author & institution disambiguation, entity extraction, triples generation
Some notions of cost

- Headline costs – beware
- Software – free to hundreds of millions
  - Support? Think what Redhat did for Linux
  - “Genuine” AI/ML/DL software?
Some notions of cost continued

- Data quality costs
  - Very large data sets (thousands to millions) must be gathered and curated
  - Data sets must be conceptually and contextually unique – (e.g. 20 to 40 for each semantic node)
  - Corrupt and inconsistent data needs normalizing and cleanup
  - Remove biased data
  - Format consistency
Some notions of cost continued

- Training costs – people, time
  - Look for software that makes training possible inhouse
    - Again, think what Redhat did for Linux
    - Need a good user interface for training tasks, app maintenance
  - US$.03 to .15 per piece at outsourcing services, but up to $2,000 for tagging a medical image, for example
  - Staff size – (Facebook – 20k and growing!)

- The cost of change – more retraining costs
Training costs case study

- Objectives for this project
  - Improve productivity
  - Improve discovery

- Goals
  - Lower cost per item
  - Discovery improved to 85% or better for recall and precision

- Process was to automatically cluster the content using NLP only vs. semantically enrich the content
Training costs case study continued

- **NLP type AI system**
  - 7500 semantic nodes to train
  - 7500 labor hours to curate training sets
    - 20 items per node needed
    - Review 60 automatically generated items to get to 20 “unique”
    - Retrain is still 1 hour per node

- **Hybrid NLP/rule layer with curated taxonomy**
  - 7500 semantic nodes to train
  - 125 labor hours to curate automatically generated rule layer
  - Retrain is <5 minutes per node
And, finally…

- Like MD Anderson, know when to cut your loses
- Good luck!