AI In Scholarly Research

Charleston Conference
23 September 2019

Haris Dindo, Jason Chabak, Ruth Pickering
Moderated by Matthew Ismail
Agenda

• AI Overview – the big picture
• The challenge: how can AI help with scholarly research?
• What is a Knowledge Graph
• How does AI-based concept search differ from traditional keyword searching?
• How does more task-specific AI differ from general AI?
• The market - when will AI-based search become ubiquitous and what will be the effect of that?
• Q&A
What is AI?
AI is now expected to be the main driver of productivity and economic growth for at least the next 20 years
The Problem...
From Data to Knowledge

Machine Learning + Computational Linguistics + Graph Theory

Computational linguistics analysis is undertaken on raw data to create a vast, multi-layered, multi-dimensional set of concepts.

Data ingestion

(Data) Neural Nets

(Dynamic) Topic Models

Stochastic Learning

Concepts are projected onto a hierarchical knowledge network that represents their relations.

Knowledge extraction

Data-driven analysis is performed against the knowledge network to detect emerging phenomena.
Artificial Intelligence (AI), Computational Linguistics (CL) and Machine Learning (ML)

The technology we use combines different techniques from the field of Artificial Intelligence, namely Machine Learning, Computational Linguistics and Knowledge Representation which results in a coherent method of extracting Knowledge from data and representing it in a format suitable for subsequent querying and exploration:

• We adopt *Deep Learning* and *Statistical Approaches*, typical of *Machine Learning*, to identify entities (known as concepts)

• We train our models based on insights from *Linguistics* and *Semantics* in order to exploit syntactical regularities in unstructured text and detect relationships between entities

• We represent entities and their relationships using an *Ontology*, a formalism typical of *Artificial Intelligence* approaches, enabling a powerful interface for Knowledge Exploration as well as Inferencing - aimed at extracting unknown relationships from the data
Concept based Approach

Conceptualization of Information to Produce Knowledge
Concepts – not Keywords

Yewno hunts for concepts, not keywords, and identifies these as objects that carry a description and a significance.

A concept is therefore not identified only by its label.

- What is a Jaguar?
- What is Quantitative Easing?
- What is Global Warming?
- ...
What is a Concept?

A concept is an atomic unit of information, composed of:

- One or more definitions
- One or more topics

**Jaguar (animal)**

The **jaguar** is the largest cat in the Americas. The jaguar has a compact body, a broad head and powerful jaws. Its coat is normally yellow and tan, but the color can vary from reddish brown to black.

**Nature / Zoology**

**Nature / Wildlife**

**Jaguar (automobile)**

**Jaguar** is the luxury vehicle brand of Jaguar Land Rover, a British multinational car manufacturer with its headquarters in Whitley, Coventry, England. Jaguar has had major success in sports car racing, particularly in the Le Mans 24 Hours.

**Engineering / Car**

**Marketing / Brands**
Quantitative easing (QE) is a type of monetary policy used by central banks to stimulate the economy when standard monetary policy has become ineffective. Quantitative easing, and monetary policy in general, can only be carried out if the central bank controls the currency used in the country.

Monetary policy is the process by which the monetary authority of a country controls the supply of money, often targeting an inflation rate or interest rate to ensure price stability and general trust in the currency.
NLP vs Computational Linguistics (1 of 2)

Distributional hypothesis: words with similar meanings occur in similar contexts
• There is a correlation between distributional similarity and meaning similarity
• This can be captured via the syntagmatic associations – words that frequently occur together should have similar meanings

This approach alone falls short with:
• Polysemy: the coexistence of many possible meanings for a word or phrase (e.g., jaguar, java)
• Synonymy: different terms with equivalent meanings (e.g., “Valid” = authorized, legitimate, licit)

We need to take the meaning of concepts into account, not just their relative position: semantic space
The distributional hypothesis is a rule-based approach, operating on the keyword level

- It does not take into account multiple meanings for the same word
- How can we deal with ambiguous concepts, such as jaguar or apple?

To overcome this difficulty, Yewno embeds the multitude of concepts into a semantic space whereby semantically related concepts are closely grouped together.
Semantic Space – a powerful, unsupervised disambiguation process

What does “jaguar” mean?

Jaguar is the luxury vehicle brand of Jaguar Land Rover, a British multinational car manufacturer with its headquarters in Whitley, Coventry, England. Jaguar has had major success in sports car racing, particularly in the Le Mans 24 Hours.

The jaguar is the largest cat in the Americas. The jaguar has a compact body, a broad head and powerful jaws. Its coat is normally yellow and tan, but the color can vary from reddish brown to black.
How does more task-specific AI differ from general AI?

**Types of Artificial Intelligence**

**Type #1: Artificial Narrow Intelligence (ANI)**

Example: RankBrain by Google and Siri by Apple

When an AI’s ability to mimic human intelligence and/or behaviour is isolated to a narrow range of parameters and contexts, it’s called ANI (also known as Weak AI or Narrow AI). All existing AI are ANI.

It’s important to keep in mind that we are talking about narrow intelligence, not low intelligence.

**Type #2: Artificial General Intelligence (AGI)**

When an AI’s ability to mimic human intelligence and/or behaviour is indistinguishable from that of a human, it’s called AGI (also known as Strong AI or Deep AI).

Most experts believe AGI is possible; however, seeing as the Fujitsu-built K, one of the world’s fastest supercomputers, took 40 minutes to simulate a single second of neural activity, I wouldn’t hold my breath.

**Type #3: Artificial Super Intelligence (ASI)**

When an AI doesn’t mimic human intelligence and/or behaviour but surpasses it, it’s called ASI.

ASI is something we can only speculate about. It would surpass all humans at all things: maths, writing books about Orks & Hobbits, prescribing medicine and much, much more. Even optimistic experts believe AGI, let alone ASI, requires decades more research, perhaps even centuries.
USE CASE 1: DEVELOPING RESEARCH SKILLS

UNDERGRADUATE STUDENTS

- Understand the big picture and how things connect together
- Drill down into individual areas of interest
- How to shape good research questions
- Introduction into more scholarly resources
USE CASE 2: INTERDISCIPLINARY RESEARCH

• Finding indirect and non-obvious relationships between concepts
• Helping identify concepts and connections where there is not yet an overload of scholarly publishing
USE CASE 3: SERENDIPITOUS DISCOVERY

ANY LEARNER OR RESEARCHER

- Discovering relevant things you didn’t know you didn’t know
**DEFINITIONS**

A combined vaccine used to prevent MEASLES; MUMPS; and RUBELLA.

The **MMR vaccine** (also known as the MPR vaccine after the Latin names of the diseases) is an immunization vaccine against measles, mumps, and rubella (German measles). It is a mixture of live attenuated vaccines of three different viruses: measles virus, mumps virus, and rubella virus.
Linking internal content to a wider data set

- Owned/created content can be visualized in the graph – to understand your network and community’s research output and knowledge.

- Every item is clearly indexed and searchable in multiple different ways: topics, sub topics, concepts, title, author etc...

- Users can find the exact sections of the documents relevant to them.

- Connections can also be found to other related items, either directly or through semantic relevance.
UNDERSTAND YOUR COLLECTIONS

Total Collections

1,134,572

15,288 added since 2017

Collection Statistic by Topics

Technology and engineering

Computer

Crafts and hobbies

Psychology

Business

Veterinary medicine

History

Foreign language

Total Collections Change by Year

1800 1900 2000 2010 2020 2030 2040 2050 2060 2070

100 200 300 400 500 600 700 800 900 1,000

Number of Collections about History Published Every Year

Social science

Health and fitness

Ecology

Humanities

Games

Mathematics

Collection Statistic by Top 50 Concepts

Yewno

Strictly confidential, not for distribution without the prior consent of Yewno
UNDERSTAND TRENDS

Users can select up to 5 topics or concepts at any time and compare trends over time.

You can compare within your own content only or across other OA content sets.

OA content sets to be selected
We all need to be prepared

- In 1965 the average term of S&P 500 company was 33 years
- By 1990 this had dropped to 20 years
- Forecast to shrink to 14 years by 2026
- c50% will be replaced over next 10 years

IN
- Netflix
- Facebook
- PayPal
- Seagate

OUT
- Kodak
- USS
- Dell
- The New York Times
AI is much more than what is often presented in the media
We have all been using AI in our day-to-day lives for years.
The Market and Challenges

2026 – Next generation search
$10 Trillion

2016 – Traditional search
$1 Trillion

Q&A

Please feel free to contact jason@yewno.com with further questions after the webinar.
Let Google SEARCH

Use Yewno to DISCOVER
So you know more.