16 PhD Dissertation Topics on Information Extraction

Mohit Bansal (UNC & Amazon Scholar)
Dan Bikel (Google)
Luna Dong (Amazon)
Eduard Hovy (CMU & DoD)
Heng Ji (UIUC & Amazon Scholar)
Dan Roth (UPenn & Amazon Scholar)
Eduard Hovy (CMU & DOD)

- Keeping IE results in the (neural) IE model
- Extracting propositional semantics from images+text
- IE with real-world semantics
3 thesis topics in IE

Eduard Hovy
CMU and DARPA
1. Keeping IE results in the (neural) IE model

- The output of an IE engine is an instantiated relation or frame. This goes to a user or a KB.

- Problems:
  - The IE engine doesn’t change after what it has just read
  - The KB is static (just additive and doesn’t generalize)
  - The user cannot change many IE parameters in a feedback/repeat cycle; has to re-train the engine

- Idea: Store IE output ‘inside’ the DNN extractor!
  - It never forgets a fact it has read
  - It updates its new reading output based on what it has read
  - It even generalizes past knowledge based on new reading

- Thesis topic: Self-extending reading+learning. Create a joint neural model for IE+KB
  - Architecture: one deep NN or two?
  - How to formally define the structure/frame to be learned and extracted?
  - How to define the formal filler types? Is it necessary?
  - How can a user access/retrieve the results?
  - How can a user engine change IE parameters for re-extraction?
  - How can extracted frames be assembled, composed into bigger ones, and re-stored?
2. Extracting propositional semantics from images+text

- Images add a lot of info to text

- **Thesis topic: joint vision+text semantics extraction**
  - How to identify the **author**? Read and interpret logo (icons, names)
  - How to identify epistemic values? Recognize and interpret visual negators (icons, text) plus the background semantics of what is being negated. False claims and their refutations may be overlaid on Informational Graphics
  - How to extract **timeline** info? (title, axis numbering, axis headings, values)
  - How to integrate results extracted from images and text? What result?

---

**Myth Unmasked**

- [FactCheck](https://factcheck.org/no-evidence-flu-vaccines-make-children-more-vulnerable-coronavirus-infections)

---

- [Measles Cases by Date of Onset, October 15, 1970—January 16, 1971](https://twitter.com/fema/status/1253801441598353412)
3. IE with real-world semantics

The IE we do today typically has 10–200 simple types/classes, nicely defined in a simple ontology.

But the semantics of the real world is messy and continuous. No realistic large ontologies are used in IE.

Thesis topic: Doing IE with real-world semantics

Use [an extract of] Wikidata as the ontology
- DWD is messy, changing, and not a clean taxonomy

Use a continuous semantic distance function. Candidates:
- Embedding-based
- Hierarchy-based
- Comparison-based (tf.idf of associated words)
- Combination
- Machine learned, using WSD353 and others

Use semantic composition of the IE output
- Extract and ‘collapse’ different representations with ‘identical’ semantics:
- Accept various frames as output
- How to score composed parts? e.g., prevent = not(cause)

Masks prevent Covid ≡ Wearing a mask causes the wearer to not be infected with Covid

| ID=E1 | type=event.prevent | arg0=DWD.mask | arg1=DWD.covid |
| ID=E2 | type=event.wear | arg0=?var.X1 | arg1=DWD.mask |
| ID=E3 | type=event.statechange.become-ill | arg0=?var.X1 | arg1=DWD.covid |
| ID=R1 | type=relation.cause | arg0=E2 | arg1=E3 |

DWD scores:
- mask—mask => 1.0
- covid—covid => 1.0
- prevention—cause => 0.63
- prevention—negation => 0.51

Assessor:
- Matches all essential aspects => 1.0
- Many parts match but not all => 0.8
- A small amount of match only => 0.3
- No match or I don’t understand it => 0

N = ½ number of Qnodes counting both sides

Final score:
= 1/N x assessor x sum(DWD scores for the sets)
= 1/4 x 1.0 x 3.14
= 0.78
Fact or Fiction: understanding and targeted summarization of long texts

A broader notion of Grounding-Quantitative Reasoning from Natural Language Text

Quantitative Reasoning from Natural Language Text
Reasoning about Quantities

- Extraction, contextualizing, scoping
  - Still open
- Reasoning about quantities
  - Mapping to Equations [Roy & Roth, 2015-2018]
- How to supervise to support this level of understanding?
  - Too many (ill-defined) latent decisions
  - End-to-End is unrealistic
- How to explain the reasoning?

She reports worsened seizure frequency, seizures now occurring up to 10/week, in clusters about 2-3 day/week. Previously reported seizures occurring about 2-3 times per month, often around the time of menses, ...

Mayor Rahm Emanuel has raised more than $15 million toward his bid for a third term – more than five times the total raised by his 10 challengers combined, campaign finance records show.

The COVID-19 pandemic in the United States is part of the .... As of October 2020, there were more than 9,000,000 cases and 230,000 COVID-19-related deaths in the U.S., representing 20% of the world's known deaths, the most deaths of any country.
Facts and Fiction

• Reading a book
  • Written from a single or multiple perspectives
  • **Follow** events/situations/people **over time**

• A range of challenges to NLU and IE
  • The novel features the character **David Copperfield**, his journey of change and growth from infancy to maturity, as many people enter and leave his life and he passes through the stages of his development. (Fiction, and you know it)
  • London and England in the 19-th century; socio-economic state, child exploitation; schools, prisons, emigration to Australia (true historical facts)

• Questions:
  • Identify Facts and Fiction
  • The Audible experience: What happened in the last 5 minutes?
  • Remind me:
    • Who is…. (but don’t spoil it for me): e.g, Steerforth
    • What does X think/feel about Y?
A broader notion of Grounding (Linking)

- A long tail of reasoning about relational information impacts accurate EDL
  - Mubarak, the wife of deposed Egyptian President Hosni Mubarak,…

- The US has now managed to upset two of its closest allies by allowing the disclosure of sensitive information
  - Who are the allies? (time dependent)

- But understanding text requires grounding many other things
  - Who is this?
  - Where is it?
  - What is happening?
  - When did it happen?
Luna Dong (Amazon)

- Few-shot learning knowledge extraction from semi-structured websites in thousands of domains
- Multi-modal knowledge extraction across texts, images, tables, and semi-structured data
SEMI-STRUCTURED DATA
MINE THE GOLD

• Industry success: >90% extraction precision

• Limitations
  - Low recall
  - Restricted scale

Xin Luna Dong, Amazon→Facebook, lunadong@gmail.com
1. DIVERSE DOMAINS

Movie

My Fair Lady
1964 · 2h 10m

Snobbish phonetics Professor Henry Higgins agrees to a wager that he can make flower girl Eliza Doolittle presentable in high society.

Director: George Cukor
Writers: Alan Jay Lerner (book of musical play) · George Bernard Shaw (from a play by)
Stars: Audrey Hepburn · Rex Harrison · Stanley Holloway

Game

Awakening is the introduction quest to WARFRAME, designed to allow players to familiarize themselves with the basics of Warframe control and combat. It was introduced with Update 29.0 (2020-08-25). Players are given a choice between various starting Warframes: Excalibur, Mag, and Volt, along with a small selection of weapons.

The Tenno: monuments of an ancient warrior caste. For generations you've slept, with no call to awake you. Until now.

---In-game description

Contents [per]
1 Cinematic Intro
2 Synopsis
3 Walkthrough
4 Notes
5 Trivia
6 Media
7 Patch History

Cinematic Intro
The intro, narrated by the Lotus, tells the tale of the Tenno. They were once an ancient warrior caste who controlled the Warframes and brought ruin to the Orekin Empire, before abruptly leaving. However, with the rise of the Grineer Empire, seeds from the ruins of The Old War that now swallow colonies whole, the Tenno begin to awaken.

Xin Luna Dong, Amazon→Facebook, lunadong@gmail.com
I. DIVERSE DOMAINS

Xin Luna Dong, Amazon à Facebook, lunadong@gmail.com

Podcast
How Your Data Powers Artificial Intelligence

Data Science Math Skills

Course

Artificial intelligence isn't something we'll see in the future. Thinking machines are already here, and nine powerful companies in the US and China control their development. The spam filter in your email inbox is AI. So are programs like Google Translate. The next level for thinking machines is when they begin learning the way humans learn. Artificial intelligence gets refined, who's keeping track of whether these machines share our motivations, desires, and hopes for the future of humanity? Amy Webb, author of “The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity,” speaks with Alexis Madrigal, a staff writer for The Atlantic who covers Silicon Valley. The views and opinions of the podcast guests are their own and do not necessarily reflect those of the Aspen Institute.

LEARN MORE
Additional Information

SPEAKERS
Amy Webb
Founder of the Future Today Institute

Alexis Madrigal
Staff writer at The Atlantic

SKILLS YOU WILL GAIN
Bayes' Theorem  Bayesian Probability  Probability  Probability Theory
I. DIVERSE DOMAINS

Dance
YOUR ULTIMATE GUIDE TO SWAN LAKE


Helgi Tomasson’s Swan Lake is part of SF Ballet’s 2021 Digital Season. It will be performed in Program 07, streaming May 20–June 9.

What Is It? Ballet’s most famous love story, featuring one of the art form’s most iconic roles. Thanks in part to this ballet’s prevalence in pop culture (think Natalie Portman in Black Swan, but also, Rudolf Nureyev on The Muppets or Taylor Swift’s ballerinas in Shake It Off), swans and ballet have become nearly synonymous in the cultural imagination. That history—plus the high drama, thrilling dancing, and gorgeous music in Helgi Tomasson’s 2009 production of Swan Lake—make it one of SF Ballet’s most popular productions.

In Short? Tragic love, a Tchaikovsky score, and swans, swans, and more swans.

Who’s It For? Fans of movies that make you cry, Olympic-level athletics, and bird-watching.

What Will I See? The story begins with a young woman named Odette . . .

Show

Show Information

Duration
2 hours and 15 minutes, including one intermission.

Audience
Recommended for ages 8 and up, but not intended for children under the age of 4. Everyone must have a ticket.

Late Seating
Latecomers are seated at the discretion of management.

Story

Frozen is the timeless tale of two sisters, pulled apart by a mysterious secret. As one young woman struggles to find her voice and harness her powers within, the other embarks on an
1. DIVERSE DOMAINS

Topic 1. Few-shot learning knowledge extraction from semi-structured websites in thousands of domains

Extracted knowledge triples

(노다지 출연 김승호 )
II. DIVERSE FORMATS OF INFORMATION

Audrey Hepburn was born as Audrey Kathleen Ruston on May 4, 1929 in Ixelles, Brussels, Belgium. Her mother, Baroness Ella Van Heemstra, was a Dutch noblewoman, while her father, Joseph Victor Anthony Ruston, was born in Úzice, Bohemia, to English and Austrian parents. After her parents' divorce, Audrey went to London with her mother where she went... See full bio »

Born: May 4, 1929 in Ixelles, Brussels, Belgium
Died: January 20, 1993 (age 63) in Tolochemaz, Vaud, Switzerland

Xin Luna Dong, Amazon→Facebook, lunadong@gmail.com
II. DIVERSE FORMATS OF INFORMATION

Xin Luna Dong, Amazon→Facebook, lunadong@gmail.com
II. DIVERSE FORMATS OF INFORMATION

Topic 2. Multi-modal knowledge extraction across texts, images, tables, and semi-structured data

Xin Luna Dong, Amazon→Facebook, lunadong@gmail.com
• **Research question 1.** Can we pre-train over vast number of sources over diverse domains on different types of information?

• **Research question 2.** Shall we do transfer learning or one-size-fits all and how many models do we need to model all kinds of diversity?

• **Research question 3.** How much can different modals leverage each other in multi-modal modeling and in co-training?

Xin Luna Dong, Amazon→Facebook, lunadong@gmail.com
Thesis topic 1: Incremental approaches to entity understanding and discovery

- Large language models such as BERT, GPT-[23] and T5 are very expensive to train, and yet the information that we would like our models to take advantage of is constantly changing. This thesis explores ways to deal with incrementally arriving data that pertains to both existing and new entities of interest. We explore methods of fine-tuning large LM’s in an efficient way, as well as methods of discovering and finding facts about new entities as they arrive in both natural language (unstructured) text and structured knowledge bases.

Thesis topic 2: Integration of structured KB’s with large language models

- Recent work has shown the promise of building large language models that have access to external knowledge, which is to say, structured knowledge that is independent of a language model’s parameters. This thesis explores novel methods for integrating external knowledge into a large LM/NLU model.

Bonus Thesis topic 3 [related to topic 2]: On the synthesis of retrieval and structured queries

- Building dense representations of entities has been shown to be an accurate and scalable way of doing not only entity linking via retrieval. Similarly, other work has shown how it is possible to build large-scale neural network approaches to encoding structured queries. This Ph.D. thesis explores novel approaches to integrating methods for representation learning and retrieval with methods that can encode structured KB queries. In particular, we hope to do better at queries involving types or quantification, where only some of the entities can be reliably obtained from a KB using structured queries, and others can be obtained only through better retrieval from unstructured corpora.
Mohit Bansal (UNC & Amazon Scholar)

- Dynamic Spatio-Temporal Knowledge Extraction and Reasoning on Videos
- Multi-Hop Reasoning on Documents and Multimodal Information
Combining symbolic semantics and distributional semantics for Information Extraction

Schema-guided generation for Cross-media cross-lingual Information Extraction and Prediction

Cross-media cross-lingual joint knowledge and text embedding representation