Interdisciplinary NLP, Data Science, Linguistics & Biomedical Informatics: My Research Journey & Tips For Getting Started

Tyler Osborne
Stony Brook University CS PhD Student, Expected 2028
Boston College CSOM & MCAS '23

Agenda

I. Cognitive States: Deep Learning Inference of Belief State in Natural Language

II. Patient-Centric Biomedical Informatics: AI Collaboration with Stony Brook Hospital

III. My Journey: How I Got Into Research, General Advice & Tips for Getting Started

IV. Open Q&A

Cognitive States: Deep Learning Inference of Belief State in Natural Language

Who Cares?

- When we speak, we convey information, but not all of that information is objective
- Often, what we convey is wrapped up in a *belief*
    - "John said Mary is coming to dinner."
    - To what degree does John believe in the factuality of his utterance?

Who Cares?

- "John said Mary is coming to dinner."
- To us, it is abundantly clear that John fully believes that his utterance is true; we want AI to have the same ability
- "John guessed that Mary may come to dinner."

Who Cares?

- This sort of analysis brings us closer to capturing the full *private state* or *cognitive state* of someone in a text
    - Set of sentiments & beliefs towards what they say
- For our purposes, the people are *sources*, the beliefs they express *targets*, and the degree to which the source believes in the factuality of their utterance, the *label*

Meet the Team

- Dr. Amittai Aviram, PhD, BC Dept. of CS
  - Prof. Aviram is my honors thesis advisor and brought me onto the project
- Principal Investigator: Dr. Owen Rambow, PhD, Stony Brook Dept. of Linguistics
  - Advises graduate students who are involved on the project
- Lead Graduate Student: John Murzaku

Language Understanding (LU) Corpus

- Corpus: Collection of text

- LU is an annotated corpus
  - ○ Humans have noted source-target pairs in the text and assigned each one a label
  - ○ The author of a sentence itself is the default source

LU Corpus

- LU's labels are:
  - *CB* for committed belief
    - "I am certain that..."
  - *NCB* for non-committed belief
    - "I am not sure but think that..."
    - "I hope that..."
  - *NA* for not applicable
    - No belief expressed

LU Corpus

"He did not speak to reporters in Jordan, but he told the Associated Press before leaving the United States that he hopes to 'separate the humanitarian work from the political issues.'"

Issues with LU

- LU is not a large corpus (<7000 english words)

- Other corpora with source-target-label annotations exist, but combining them natively is next to impossible
  - Why?

Issues with LU

- However, if we could somehow port each individual corpus into a single, unified format, then we could combine them!
    - This was the basis of my honors thesis

Factbank: A Natively Relational Corpus

- Factbank is another belief state annotation corpus falling under the source-target-label paradigm
  - Different label scheme

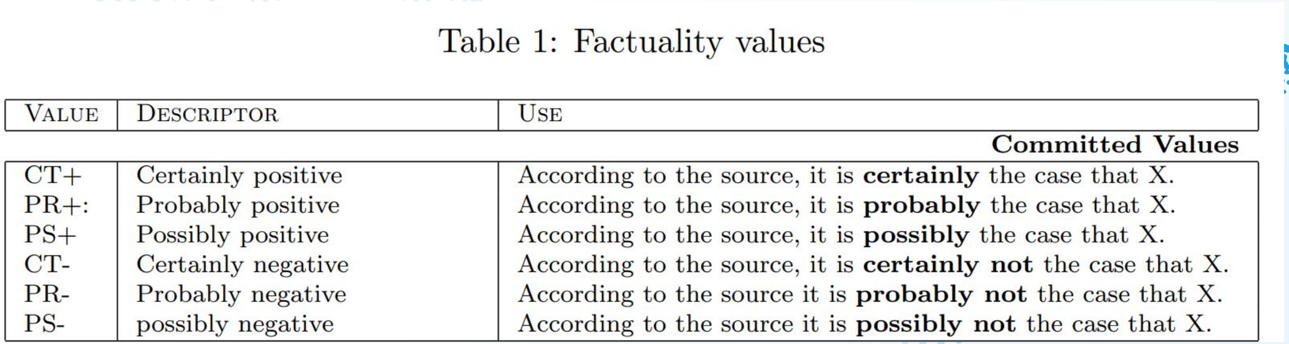


Table 1: Factuality values

VALUE	DESCRIPTOR	USE
		Committed Values
CT+	Certainly positive	According to the source, it is **certainly** the case that X.
PR+:	Probably positive	According to the source, it is **probably** the case that X.
PS+	Possibly positive	According to the source, it is **possibly** the case that X.
CT-	Certainly negative	According to the source, it is **certainly not** the case that X.
PR-	Probably negative	According to the source it is **probably not** the case that X.
PS-	possibly negative	According to the source it is **possibly not** the case that X.

Factbank: A Natively Relational Corpus

Sentence	Target Head	Source Text	Label
… for an economy that many experts thought was once invincible.	invincible	Author	CT+
… for an economy that many *experts* thought was once invincible.	invincible	experts_Author	CT+

LU vs. Factbank

LU

- Three labels
- Bona fide flat files (XML)
- Author-only annotations

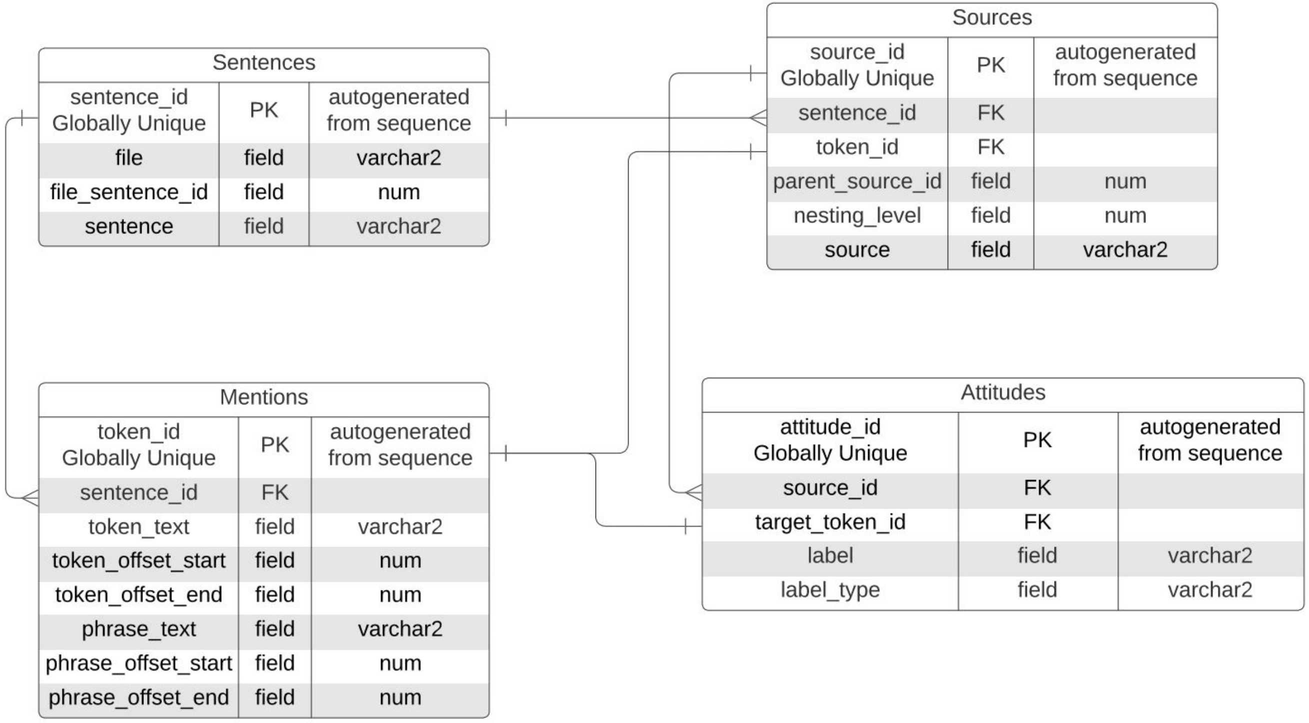
Factbank

- Six labels
- Relational data stored as flat files
- Author & nested source annotations

Conclusion:

Factbank much more complex;
impossible to (natively) combine!

Unified Database Model: Entity-Relation Diagram



Sentences

sentence_id Globally Unique	PK	autogenerated from sequence
file	field	varchar2
file_sentence_id	field	num
sentence	field	varchar2

Sources

source_id Globally Unique	PK	autogenerated from sequence
sentence_id	FK	
token_id	FK	
parent_source_id	field	num
nesting_level	field	num
source	field	varchar2

Mentions

token_id Globally Unique	PK	autogenerated from sequence
sentence_id	FK	
token_text	field	varchar2
token_offset_start	field	num
token_offset_end	field	num
phrase_text	field	varchar2
phrase_offset_start	field	num
phrase_offset_end	field	num

Attitudes

attitude_id Globally Unique	PK	autogenerated from sequence
source_id	FK	
target_token_id	FK	
label	field	varchar2
label_type	field	varchar2

Unified DB Model: Data Transformations

- **Goal:** Preserve native data while inserting synthetic data where gaps appear
    - Ex: MPQA has a reported belief class, Factbank does not

Unified DB Model: Data Transformations

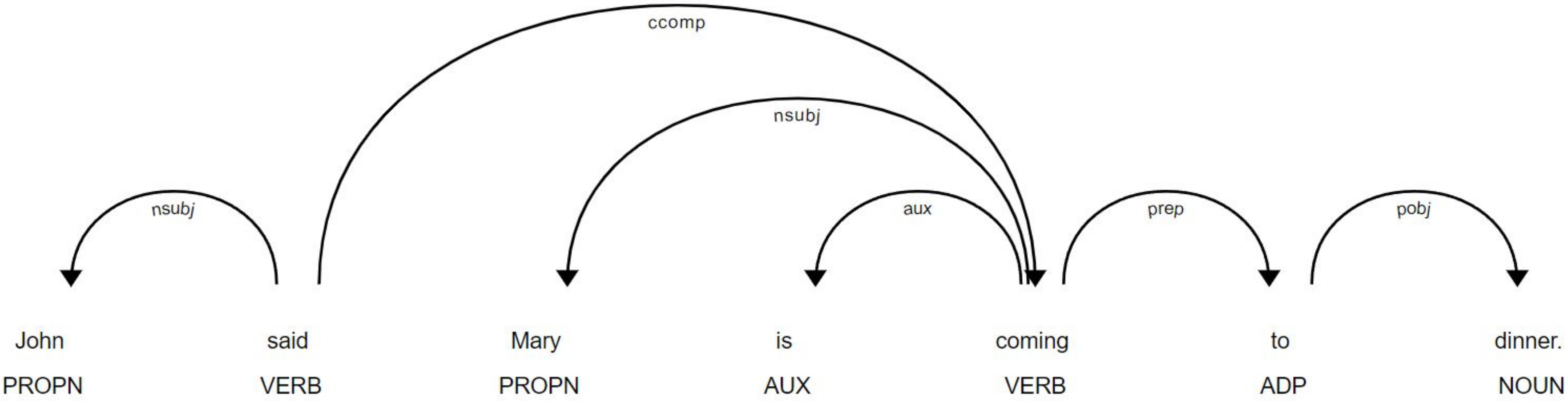
I. Unigram heads ↔ N-gram spans

    A. Parse trees! 🌳

II. Additional Classes (Factbank: ROB, LU: O)

Quick Aside: Dependency Parse Trees

- `spaCy` library in Python (`displayCy` 😁)



Unigram Heads ↔ N-Gram Spans

- **Goal:** Extract embedded proposition containing the target (noun or verb phrase)
    - Parse trees contain noun/verb phrases

- Factbank target head words live inside one of these phrases or may head it

fb2master.py

```
def get_span:
    if head_token is ROOT:
        return head_token
    for each token in head_token's ancestors:
        if token in [PRON, PROPN, NOUN, VERB, AUX]:
            return (token.left_edge, token.right_edge)
```

Additional Classes

- Factbank: Reported Belief (ROB)
  - Natively grouped with Uu

- LU: Other (O)
  - Natively unannotated

fb2master.py

- The porting task required more nuanced logic than simply iterating over a result set
  - Trust me, we tried the easy way

- We needed to design... wait for it...
  - An algorithm 😎

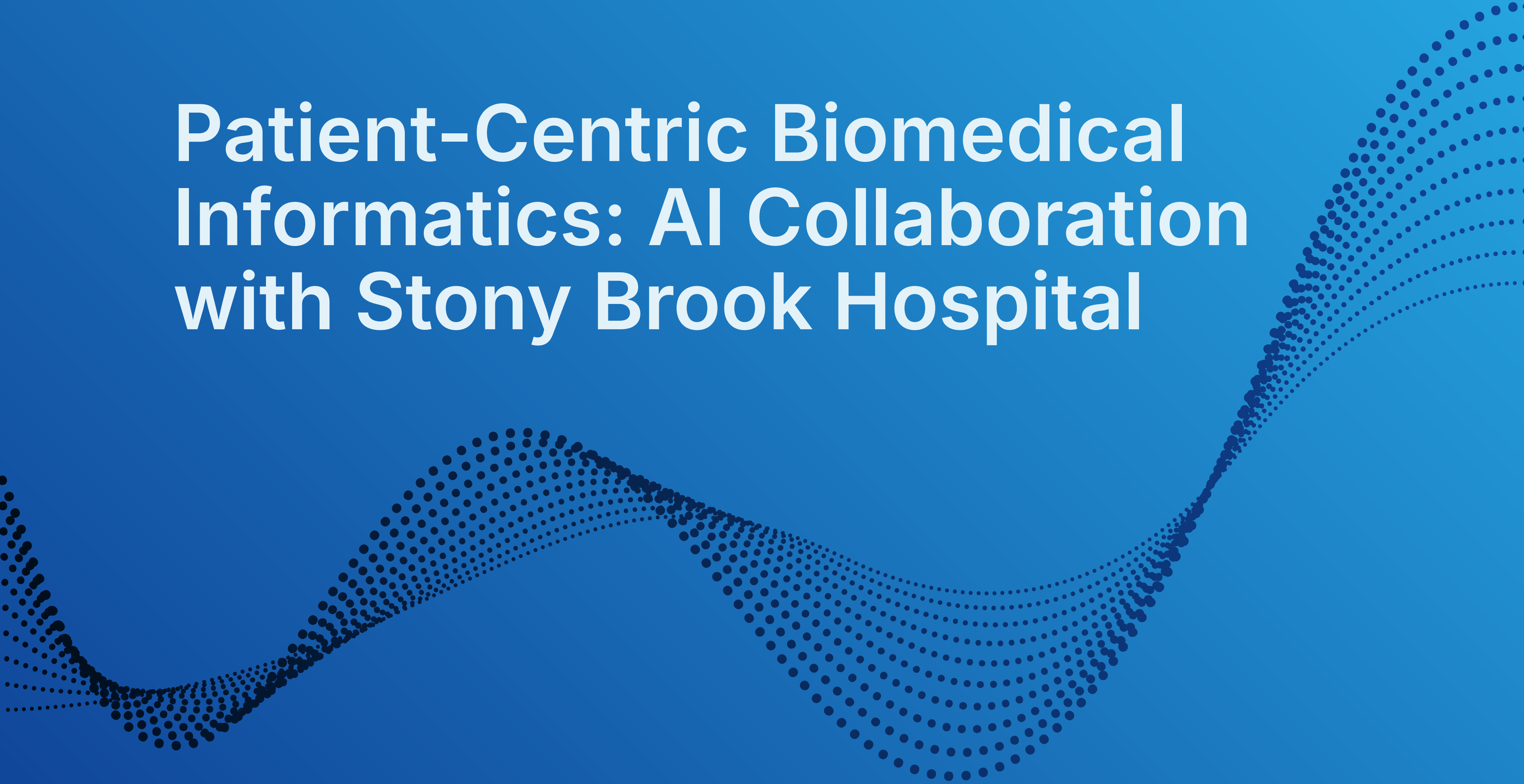
fb2master.py

```
def fb2master:
    for each sentence:
    sources ← all sources from that sentence
    for the sources on each nesting level, 0 to 3:
        catalog source in mentions table
        parent_source ← source's parent (if exists)
        catalog source with parent_source in sources table
        for each target relevant to this source:
            catalog target in mentions table
            catalog label corresponding to source/target, in attitudes table
```

`fb2master.py`

- That's a lot of `for` loops... how to optimize?
  - Reducing serialization: SQLite overhead
  - Runtime reduction from 5 mins to 3 seconds!
- Aside: Do we care about efficiency in this case?
  - Yes & no...

Patient-Centric Biomedical Informatics: AI Collaboration with Stony Brook Hospital



Patient-Centric Biomedical Informatics

- AI can automate many tasks in medicine, theoretically freeing up doctors to focus on patient care directly
- How do we ensure medical AI is trustworthy, while tailoring it to individual patient preferences and beliefs?

AI-Generated Discharge Summaries

- Documentation Burden → Clinician Burnout
- LLMs are really good at summarizing!
- Informing CS methodologies via MD expertise
- Data [un]availability

My Journey: How I Got Into Research, General Advice & Tips for Getting Started

My Journey

- It all started with the CS TA program

- Prof. Aviram agreed to advise an independent study for the Cognitive States project, which turned into my honors thesis (9 CS elective credits!)

My Journey

- Being involved in research made me realize I was not looking forward to the industry track
- I applied to PhD programs my senior year
- At Stony Brook, I wanted to apply NLP to a higher-level task in a field where my work would be genuinely helpful to society

Tips For Getting Started

- Do well in your courses and TA for them later on – **Ask the professor to make custom arrangements for you to TA for them**

    - While getting to know your professors is a good idea in general, the TA program makes it much easier and more natural

Tips For Getting Started

- Look up the Google Scholar profiles of professors you might want to work with and check out their work

- Nobody, including me, was good at doing research when they started
    - Be okay with things not working out

Conclusion

- This kind of work pushes you to grow as a computer scientist in ways courses cannot
  - I had **NO CLUE** what I was doing when I started on this project

Open Q&A