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

Tyler Osborne, CSOM & MCAS '23

Part I: Who Cares?

Why is this an important topic in NLP?

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*

Part II: Meet the Team

Who are the key people involved in this work?

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 some of the ~10 graduate students who are involved on the project

Part III: Language Understanding (LU) Corpus

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 huge 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!
  - Prof. Aviram and I have spent the past year on this task

Part III: Factbank & Unified Database Model

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

- Factbank is stored relationally
  - Makes most sense as a database

- Porting the raw Factbank files into a database was easy
  - One SQL table per file, one row in SQL table per line in file

Factbank: A Natively Relational Corpus

```
'ABC19980108.1830.0711.tml'|||0|||'ABC19980108.1830.0711'
'ABC19980108.1830.0711.tml'|||1|||'On the other hand, it\'s turning out to be another very bad financial week for Asia.'
'ABC19980108.1830.0711.tml'|||2|||'The financial assistance from the World Bank and the International Monetary Fund are not helping.'
'ABC19980108.1830.0711.tml'|||3|||'In the last twenty four hours, the value of the Indonesian stock market has fallen by twelve percent.'
'ABC19980108.1830.0711.tml'|||4|||'The Indonesian currency has lost twenty six percent of its value.'
'ABC19980108.1830.0711.tml'|||5|||'In Singapore, stocks hit a five year low.'
'ABC19980108.1830.0711.tml'|||6|||'In the Philippines, a four year low.'
'ABC19980108.1830.0711.tml'|||7|||'And in Hong Kong, a three percent drop.'
'ABC19980108.1830.0711.tml'|||8|||'More problems in Hong Kong for a place, for an economy, that many experts thought was once invincible.'
'ABC19980108.1830.0711.tml'|||9|||'Here\'s ABC\'s Jim Laurie.'
'ABC19980108.1830.0711.tml'|||10|||'Not that long ago, before the Chinese takeover, the news about real estate here was that the sky was the limit the highest pr
'ABC19980108.1830.0711.tml'|||11|||'So when Wong Kwan spent seventy million dollars for this house, he thought it was a great deal.'
'ABC19980108.1830.0711.tml'|||12|||'He sold the property to five buyers and said he\'d double his money.'
'ABC19980108.1830.0711.tml'|||13|||'In Hong Kong, is always belongs to the seller\'s market.'
'ABC19980108.1830.0711.tml'|||14|||'Now with new construction under way, three of his buyers have backed out.'
'ABC19980108.1830.0711.tml'|||15|||'And Wong Kwan will be lucky to break even.'
'ABC19980108.1830.0711.tml'|||16|||'All across Hong Kong, the property market has crashed.'
'ABC19980108.1830.0711.tml'|||17|||'Pamela Pak owns eight condominiums here.'
```

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

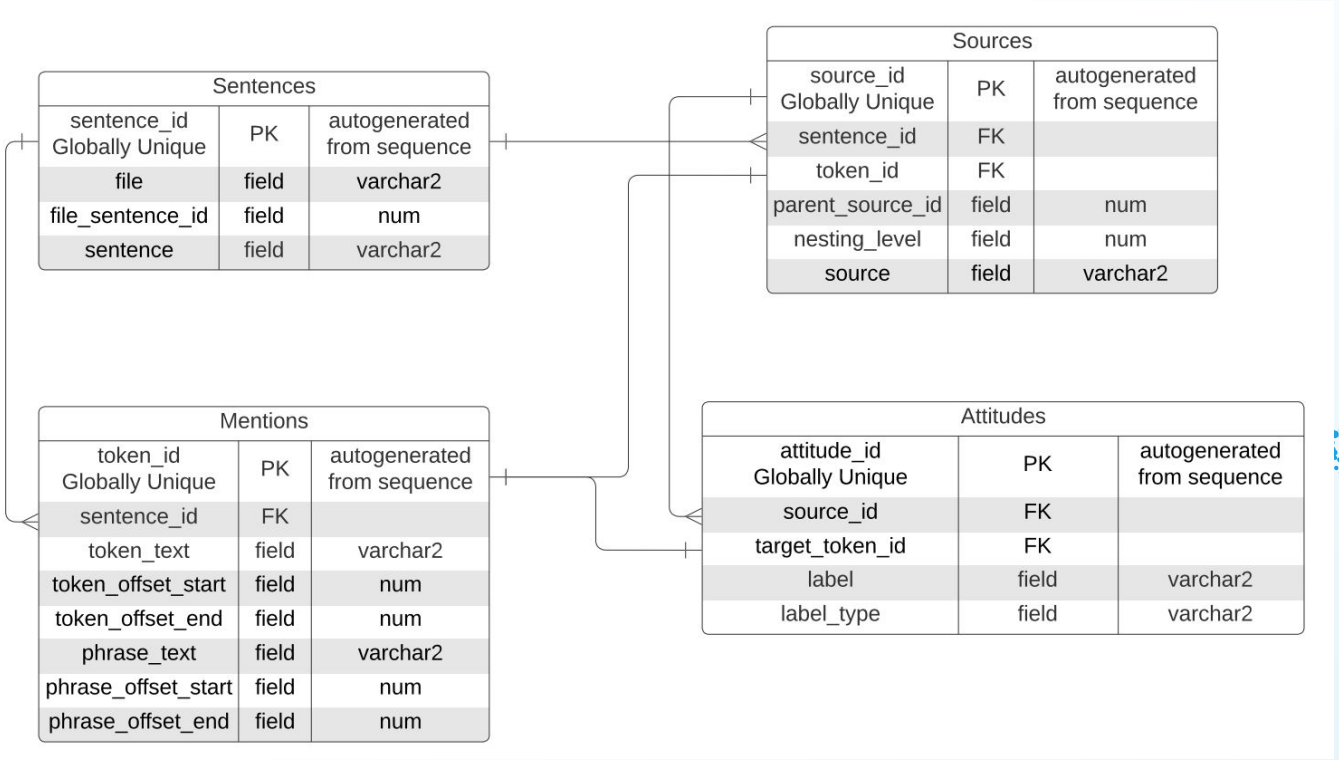
- Three labels
- Bona fide flat files (XML)
- Author-only annotations

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



Part IV: `fb2master.py`

The meat & potatoes of my work

`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...

`fb2master.py`

- Months of debugging later: an accurate, clean & intuitive database representation of Factbank

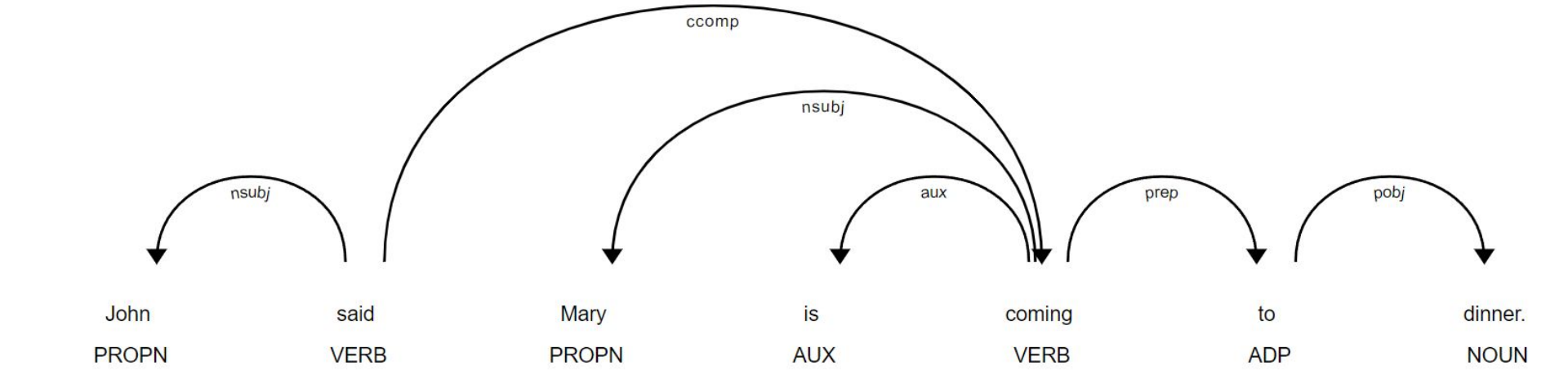
- Wait, there's more!

`fb2master.py`

- Bells & Whistles
  - Reported belief class (ROB)
    - Another algorithm, will not cover
  - Source/target spans
    - Brought to you by the `spaCy` library!

fb2master.py

- spaCy library: parse trees



fb2master.py

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

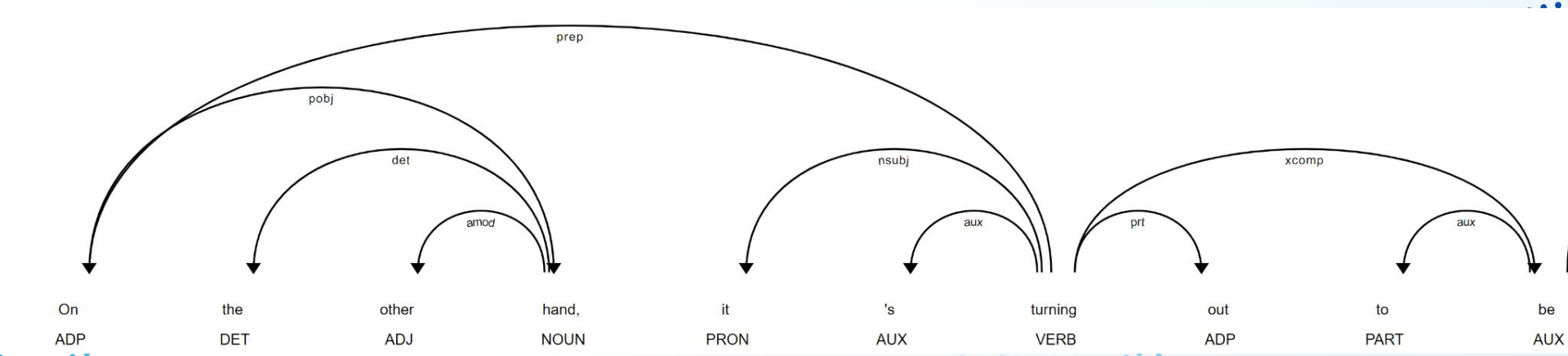
- Factbank target head words live inside one of these phrases or may head it

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)
```

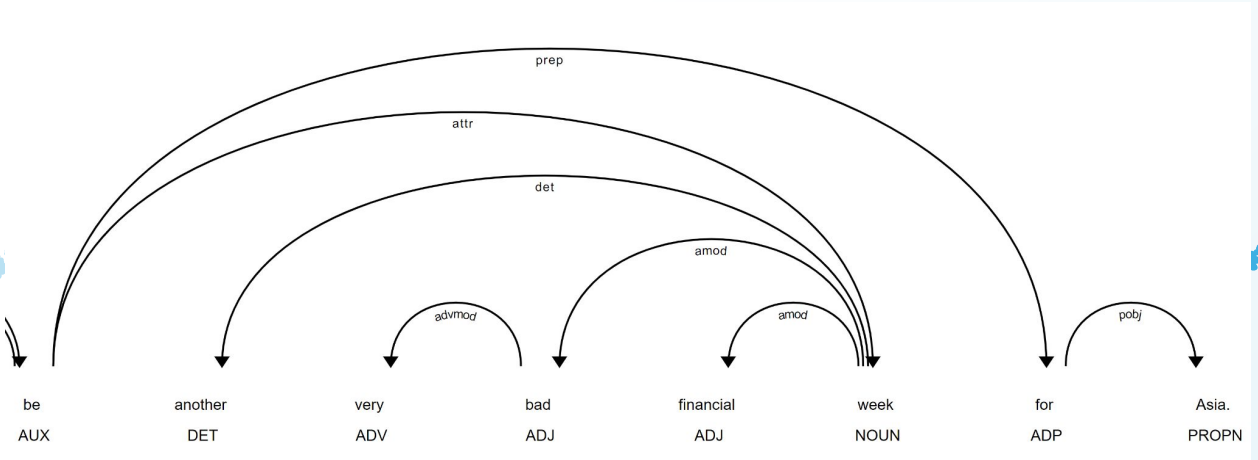
fb2master.py

- Example: "On the other hand, it's turning out to be another very **bad** financial week for Asia."



fb2master.py

- Example: "On the other hand, it's turning out to be another very **bad** financial week for Asia."



`fb2master.py`

- Script runtime after `spaCy` spans implementation: ~30 seconds (AMD Ryzen 5 6-Core, 4.2GHz)
    - Why 10x slower?

Part V: Looking Ahead & Conclusions

What's next for this research?

Looking Ahead

- Import more corpora annotated for belief
  - BeSt, MPQA, etc

- Deep learning experiments
  - Compare performance between input data formats
  - Could we beat SOTA?

Conclusions

- 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
- You can do research for course credit
  - 9 CSCI4000 credits = 3 electives
  - Get to know your professors
  - Happy to discuss & give tips offline!