Text as Data: Semantic Parsing

Guest Course – January 2026

Germain Gauthier, Philine Widmer[1]

[1]Bocconi Unversity, Paris School of Economics

USI Lugano

So far: representations and patterns

- Previous sessions focused on **representations**:
  - Bag-of-words: Documents as word counts
  - Topic models: Documents as topic mixtures
  - Embeddings: Words as dense vectors
- These methods capture **statistical patterns** but often ignore **linguistic structure**
- **This session**: Extract structured information from text
  - Who did what to whom? (Semantic Role Labeling)
  - What entities are mentioned? (Named Entity Recognition)
  - How do words relate grammatically? (Dependency Parsing)
  - Who do pronouns refer to? (Co-reference Resolution)

Part-of-Speech (PoS) Tagging

Goal: Label each word with its grammatical category

Example:

"The Federal Reserve raises interest rates."

- The: `DET` (determiner)
- Federal Reserve: `PNOUN` (proper noun)
- raises: `VERB` (verb)
- interest: `NOUN` (noun)

Why useful?

- Text cleaning (e.g., only keep nouns, adjectives, verbs)
- Disambiguation ("interest" as noun vs. verb)

Named Entity Recognition (NER)

Goal: Identify and classify named entities in text

Common entity types:

- `PERSON`: Barack Obama, Janet Yellen
- `ORGANIZATION`: Federal Reserve, Goldman Sachs, IMF
- `LOCATION`: United States, Wall Street, Europe
- `DATE`: January 2023, 9/11
- `MONEY`: $5 billion, 2.3% GDP
- `PERCENT`, `CARDINAL`, `ORDINAL`, etc.

Example:

"[The Federal Reserve]$_{ORG}$ raised rates by [0.25%]$_{PERCENT}$ in [March]$_{DATE}$."

NER: Why it matters

Applications in social science:

- **Political science**: Track politician mentions, policy topics
- **Economics**: Extract firm names, financial metrics, locations
- **Sociology**: Identify actors, institutions, demographics
- **History**: Extract historical figures, dates, places

Examples:

- Measuring media bias: Which politicians are mentioned and how?
- Corporate networks: Which firms are mentioned together?

Dependency Parsing

Goal: Identify grammatical relationships between words

Each word has a **head** (the word it depends on) and a **relation** (the type of dependency)

Example: *"The economy is growing rapidly."*

- economy $\xrightarrow{\text{det}}$ The
- growing $\xrightarrow{\text{nsubj}}$ economy
- growing $\xrightarrow{\text{aux}}$ is
- growing $\xrightarrow{\text{advmod}}$ rapidly

Universal Dependencies (UD):

- Standardized annotation scheme
- Relations: `nsubj` (subject), `obj` (object), `obl` (oblique), etc.

Dependency Parsing: Why it matters

Captures sentence structure beyond word order:

- *"The Fed raised rates"* $\neq$ *"Rates raised the Fed"*
- Same words, different dependencies!

Applications:

- **Relation extraction**: Who did what to whom?
- **Information extraction**: Extract structured facts
- **Sentiment analysis**: Identify opinion targets

Example: Extract subject-verb-object triples

- Input: *"Congress passed the bill."*
- Output: (Congress, passed, bill)

Semantic Role Labeling (SRL)

Goal: Identify "who did what to whom, when, where, why, how?"

Semantic roles:

- `ARG0`: Agent (doer)
- `ARG1`: Patient (thing affected)
- `ARG2`: Instrument, beneficiary, etc.
- `ARGM-TMP`: Temporal (when)
- `ARGM-LOC`: Location (where)
- `ARGM-MNR`: Manner (how)

Example:

"[The Fed]$_{ARG0}$ raised [rates]$_{ARG1}$ [yesterday]$_{ARGM\text{-}TMP}$."

SRL: Why it matters

Deeper than syntax:

- *"John broke the window"* vs. *"The window broke"*
- Different syntax, but "window" is ARG1 (patient) in both

Applications in social science:

- **Media framing**: How are events described? Active vs. passive voice?[1]
- **Political analysis**: Extract political and economic narratives from text[2]

Example research questions:

- How do newspapers frame protests? (Are protesters agents or patients?)
- Who gets credit/blame in political speeches?

Co-reference Resolution

Goal: Determine which mentions refer to the same entity

Example:

> *"Janet Yellen testified before Congress yesterday. The Fed Chair said she remains committed to fighting inflation. Yellen emphasized that her approach would be data-driven."*

All underlined mentions refer to the same person (Janet Yellen).

Types of mentions:

- Proper names: "Janet Yellen"
- Definite descriptions: "The Fed Chair"
- Pronouns: "she", "her"

Co-reference Resolution: Why it matters

Critical for understanding discourse:

- Who/what is being discussed across sentences?
- Track entities throughout documents
- Resolve ambiguous references

Applications:

- **Information extraction**: Link mentions to build entity profiles
- **Question answering**: Resolve pronouns in questions/answers
- **Summarization**: Avoid redundant mentions
- **Network analysis**: Build co-mention networks

Co-reference Resolution: Challenges

Key challenges:

- **Long-range dependencies**: References can span many sentences
- **Ambiguity**: "it" could refer to many entities
- **World knowledge**: "The Fed" = "The Federal Reserve"
- **Metonymy**: "Washington" referring to the U.S. government

Putting it all together: A pipeline

Typical NLP pipeline:

1. **Tokenization**: Split text into words/sentences
2. **PoS tagging**: Label grammatical categories
3. **Dependency parsing**: Identify grammatical relations
4. **NER**: Recognize named entities
5. **Co-reference**: Link mentions

Available tools:

- `spaCy`: Fast, production-ready, Python
- `AllenNLP`: Research-oriented, many pre-trained models

Summary

Semantic parsing extracts structured meaning from text:

- **PoS tagging**: Grammatical categories (foundation)
- **NER**: Identify and classify entities
- **Dependency parsing**: Grammatical relationships
- **SRL**: Who did what to whom?
- **Co-reference**: Link mentions of same entity

Key takeaways:

- Captures **linguistic structure** and **semantic relations**
- Useful for information extraction and event detection
- Modern tools make it accessible (spaCy, AllenNLP)

References I

Ash, Elliott, Germain Gauthier, and Philine Widmer (2024). “Relatio: Text semantics capture political and economic narratives”. In: *Political Analysis* 32.1, pp. 115–132.

Moreno-Medina®, Jonathan et al. (2025). “Officer-involved: The media language of police killings”. In: *The Quarterly Journal of Economics* 140.2, pp. 1525–1580.