Overview of Annotation Tasks II

Sequence Labeling Tasks

Jin Zhao

Brandeis University

February 4, 2025

Today's Agenda

1. Review of classification tasks
2. Introduction to sequence labeling
3. Named Entity Recognition (NER)
4. Part-of-Speech tagging
5. BIO/IOB tagging schemes
6. Span annotation challenges
7. Guidelines for sequence labeling

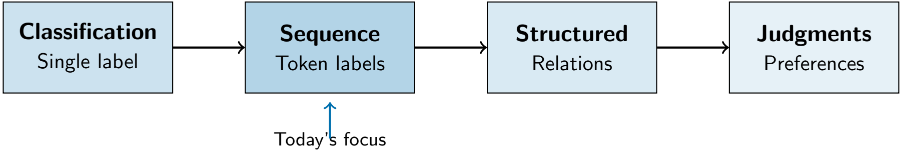
Quick Review: Classification

From last lecture:

- Classification assigns one or more labels to a text unit
- Sentiment analysis, topic classification, intent detection
- Multi-class (exclusive) vs. multi-label (non-exclusive)
- Schema design: mutually exclusive, exhaustive, clear boundaries

Today: Moving from document-level to token-level annotation

The Task Spectrum



Sequence Labeling: Assign a label to each token in a sequence

- More complex than classification
- Introduces span boundary decisions
- Moderate inter-annotator agreement

What is Sequence Labeling?

Definition: Assigning labels to individual tokens or spans within a text

Key characteristics

- Each token gets a label (including "Outside")
- Labels often form contiguous spans
- Order and context matter
- Need to handle span boundaries

Common tasks:

- Named Entity Recognition (NER)
- Part-of-Speech (POS) tagging
- Chunking (noun phrases, verb phrases)
- Slot filling in dialogue systems

Named Entity Recognition (NER)

Goal: Identify and classify named entities in text

Standard entity types:

- **PER** – Person names
- **ORG** – Organizations
- **LOC** – Locations
- **DATE** – Temporal expressions
- **MISC** – Miscellaneous entities

Example:
"[Apple]$_{ORG}$ announced a new [iPhone]$_{PROD}$ in [Cupertino]$_{LOC}$."

Challenge: "Apple" could be a company, fruit, or person's name

NER Annotation Decisions

What counts as an entity?

Clear cases:

- “Barack Obama” – PER
- “Google” – ORG
- “New York City” – LOC
- “January 15, 2025” – DATE

Difficult cases:

- “the president” – PER?
- “American” – nationality?
- “iPhone” – product?
- “COVID-19” – disease?

Guideline questions:

- Do we annotate generic mentions (“the company”)?
- How do we handle nested entities?
- What about abbreviated names?

Part-of-Speech Tagging

Goal: Assign grammatical categories to each word

Common POS tags:

- NN – Noun
- VB – Verb
- JJ – Adjective
- RB – Adverb
- DT – Determiner
- IN – Preposition
- PRP – Pronoun
- CC – Conjunction
- . – Punctuation
- CD – Cardinal number

Example:

The/DT quick/JJ brown/JJ fox/NN jumps/VBZ over/IN the/DT lazy/JJ dog/NN

Challenge: Many words have multiple possible POS (“run” can be noun or verb)

The BIO Tagging Scheme

Standard format for sequence labeling

- **B** – Beginning of an entity
- **I** – Inside (continuation) of an entity
- **O** – Outside any entity

Example:

Token	Apple	announced	iPhone	in	New	York
BIO	B-ORG	O	B-PROD	O	B-LOC	I-LOC

Why B and I? To distinguish adjacent entities of the same type
"[John]$_{PER}$ [Smith]$_{PER}$" vs. *"[John Smith]$_{PER}$"*

BIO Variants

IOB1 vs. IOB2 (BIO):

- **IOB1:** B only used when two entities are adjacent
- **IOB2 (BIO):** B always starts a new entity (most common)

BIOES/BILOU:

- **B** – Beginning
- **I** – Inside
- **O** – Outside
- **E/L** – End of entity (last token)
- **S/U** – Single-token entity

Example with BIOES:

Token	Apple	announced	iPhone	in	New	York
BIOES	S-ORG	O	S-PROD	O	B-LOC	E-LOC

Span Boundary Challenges

Where does an entity start and end?

Common difficulties

- **Modifiers:** "the [Microsoft] Corporation" or "[the Microsoft Corporation]"?
- **Titles:** "[President Biden]" or "President [Biden]"?
- **Possessives:** "[Apple's] iPhone" – is "'s" part of the entity?
- **Coordination:** "[North and South Korea]" – one or two entities?

Solution: Clear guidelines with examples for each case

- Document your boundary conventions
- Be consistent within your dataset
- Include edge cases in annotator training

Nested and Overlapping Entities

Problem: Some entities contain other entities

Example:
"The [University of [California]$_{LOC}$]$_{ORG}$"

Options:

1. **Flat annotation:** Only annotate outermost entity
2. **Nested annotation:** Allow entities to contain others
3. **Multiple passes:** Separate annotation layers

BIO limitation: Standard BIO cannot represent nested entities
Solutions:

- Extended tagging schemes (nested BIO)
- Separate annotation layers
- Different data format (standoff, JSON)

Tokenization Matters

Sequence labeling depends on tokenization

Tokenization decisions affect annotation

- "New York" – 1 token or 2?
- "don't" – 1 token ("don't") or 2 ("do", "n't")?
- "U.S.A." – how many tokens?
- Hyphenated words: "state-of-the-art"

Best practices:

1. Tokenize **before** annotation
2. Document your tokenization rules
3. Use consistent tokenization for train/test
4. Consider subword tokenization for models

Annotation Tools for Sequence Labeling

Tools designed for span annotation:

- **brat:** Classic tool, standoff format, good for NER
- **Label Studio:** Modern, flexible, supports multiple formats
- **Prodigy:** Active learning, spaCy integration
- **Doccano:** Simple, open-source

Key features to look for:

- Easy span selection (click and drag)
- Keyboard shortcuts for labels
- Pre-annotation / suggestions
- Export to BIO format

IAA for Sequence Labeling

Measuring agreement is more complex than classification

Token-level agreement:

- Treat each token as an instance
- Calculate Cohen's/Fleiss' Kappa
- Simple but doesn't capture span structure

Entity-level agreement:

- Count matching spans (exact or partial)
- Precision, recall, F1 between annotators
- Better reflects actual task

Typical targets:

- Token-level: $\kappa > 0.8$
- Entity-level F1: > 0.85 for clear tasks

Guidelines for Sequence Labeling

Essential elements:

1. **Entity definitions:** What qualifies as each type?
2. **Boundary rules:** Where entities start and end
3. **Examples:** Positive and negative for each type
4. **Edge cases:** Abbreviations, nicknames, titles
5. **Nesting policy:** How to handle overlapping entities

Example guideline entry:

PERSON (PER)

Names of people, including fictional characters. Include titles only if they are part of the proper name. Do NOT annotate generic references like "the president."

✓ "Barack Obama", "Dr. Smith", "Queen Elizabeth II"

× "the CEO", "my brother", "the author"

LLM Annotation for Sequence Labeling

More challenging than classification for LLMs

Challenges:

- Need to maintain token alignment
- Boundary decisions can be inconsistent
- Output format must be precise

Strategies:

- Request JSON output with character offsets
- Use few-shot examples with exact format
- Post-process to align with tokenization
- Validate output programmatically

Best approach: LLM pre-annotation + human correction
LLM suggests entities, humans verify and fix boundaries

Data Format: From Annotation to Training

Converting annotations to model input:

Annotation format (standoff):

- Text: "Apple is in Cupertino"
- T1: ORG 0 5 "Apple"
- T2: LOC 12 21 "Cupertino"

Training format (BIO):

- Apple B-ORG
- is O
- in O
- Cupertino B-LOC

Conversion requires:

- Tokenization of source text
- Mapping character offsets to tokens
- Handling multi-token entities
- Export scripts (Label Studio, brat provide these)

Next Class: MATTER Cycle Deep Dive

Lecture 8 (Feb 9): Task Formalization

Topics:

- Formalizing annotation tasks
- Document Type Definitions (DTDs)
- JSON Schema for annotation specifications
- Tag types: non-consuming, span, link, attribute
- Prompts as lightweight task specifications

Reading: Pustejovsky & Stubbs, Chapter 5

Key Takeaways

1. **Sequence labeling** assigns labels to each token in a text
2. **NER** identifies named entities – boundary decisions are crucial
3. **BIO scheme** is the standard format: Beginning, Inside, Outside
4. **Span boundaries** are the main source of annotator disagreement
5. **Tokenization** must be done before annotation and documented
6. **IAA** can be measured at token-level or entity-level

Questions?

Questions?

Office Hours: Wednesdays 1-3pm, Volen 109

✉ jinzhao@brandeis.edu