Natural Language Processing

Class 8: Sequence Labeling Tasks

Adam Faulkner

October 28, 2025

Why do we need sentence structure?

- Humans communicate complex ideas by composing words together into bigger units to convey complex meanings
- Listeners need to work out what modifies [attaches to] what
- A model needs to understand sentence structure in order to be able to interpret language correctly

Constituency Grammar (CFG) Review

- The traditional approach to syntax (Phrase Structure Grammar).
- Focuses on constituents (phrases) that act as single units.
- Structure is represented by a tree with non-terminal (phrase) nodes (*NP*, *VP*, *S*).

Constituency Grammar (CFG) Review

- The traditional approach to syntax (Phrase Structure Grammar).
- Focuses on constituents (phrases) that act as single units.
- Structure is represented by a tree with non-terminal (phrase) nodes (*NP*, *VP*, *S*).

The Core Limitation

CFG focuses on grouping words, often obscuring the direct relationships (who does what to whom). It doesn't explicitly model the head-dependent relationship.

Constituency Structure (phrase structure)

- Syntactic constituency is the idea that groups of words can behave as single units, or constituents. Part of developing a grammar involves building an inventory of the constituents in the language.
- Phrase structure organizes words into nested constituents.
- How do we know what a constituent is?
- Distribution: A constituent behaves as a unit that can appear in different places:
  - John talked [to the children] [about drugs].
  - John talked [about drugs] [to the children].
  - *John talked drugs to the children about

  Substitution/expansion/pronoun replacement
  - I sat [on the box/right on top of the box/there].

Chomsky, 1956: Context-Free Grammars

- A context-free grammar consists of a set of rules or productions, each of which expresses the ways that symbols of the language can be grouped and ordered together, and a lexicon of words and symbols.
- For example, the following productions NP express that an NP (or noun phrase) can be composed of either a ProperNoun or a determiner (Det) followed by a Nominal; a Nominal in turn can consist of one or more Nouns.

- NP → Det Nominal
- NP → ProperNoun
- Nominal → Noun | Nominal Noun

Chomsky, 1956: Context-Free Grammars

- Context-free rules can be hierarchically embedded, so we can combine the previous rules with others, like the following, that express facts about the lexicon:
  - Det → a
  - Det → the
  - Noun → flight
- Terminals: The symbols that correspond to words in the language (the, nightclub); the lexicon is the set of rules that introduce these terminal symbols.
- Non-terminals: In each context-free rule, the item to the right of the arrow (→) is an ordered list of one or more terminals and non-terminals; to the left of the arrow is a single non-terminal

Chomsky, 1956: Context-Free Grammars

$G = (T, N, S, R)$

- T is a set of terminal symbols
- N is a set of nonterminal symbols
- S is the start symbol ($S \in N$)
- R is a set of rules/productions of the form $X \rightarrow \gamma$
  - $X \in N$ and $\gamma \in (N \cup T)^*$

A grammar G generates a language L.

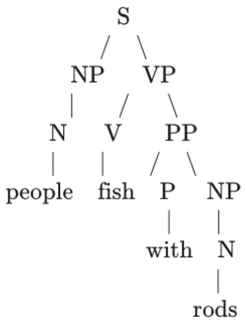
Given the following CFG...

S → NP VP
VP → V NP
VP → V NP PP
NP → NP NP
NP → NP PP
NP → N
NP → *e*
PP → P NP

N → people
N → fish
N → tanks
N → rods
V → people
V → fish
V → tanks
P → with

people fish tanks
people fish with rods

...we can generate a parse for “people fish with rods”



We generated this parse using

- S → NP VP
- VP → V PP
- PP → P NP
- NP → N
- N → people, fish, rods
- P → with

Probabilistic context-free grammars (PCFGs)

G = (T, N, S, R, P)

- T is a set of terminal symbols
- N is a set of nonterminal symbols
- S is the start symbol ($S \in N$)
- R is a set of rules/productions of the form $X \rightarrow \gamma$
- P is a probability function
  - $P: R \rightarrow [0,1]$
  - $\forall X \in N, \sum_{X \rightarrow \gamma \in R} P(X \rightarrow \gamma) = 1$

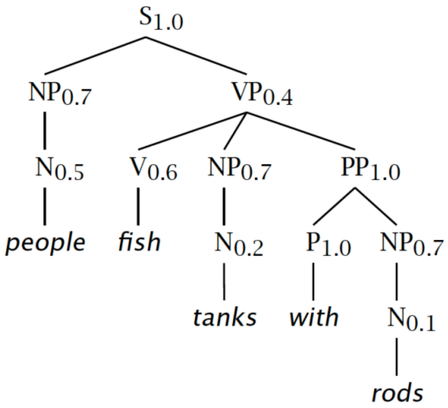
A grammar G generates a language model L.

$$\sum_{\gamma \in T*} P(\gamma) = 1$$

The probability of trees and strings

- $P(t)$ The probability of a tree t is the product of the probabilities of the rules used to generate it.
- $P(s)$ The probability of the string s is the sum of the probabilities of the trees which have that string as their yield

$$P(s) = \sum_j P(s,t), \text{ where t is a parse of s} = \sum_j P(t)$$


S1.0
NP0.7
VP0.4
N0.5
V0.6
NP0.7
PP1.0
people
fish
N0.2
P1.0
NP0.7
tanks
with
N0.1
rods

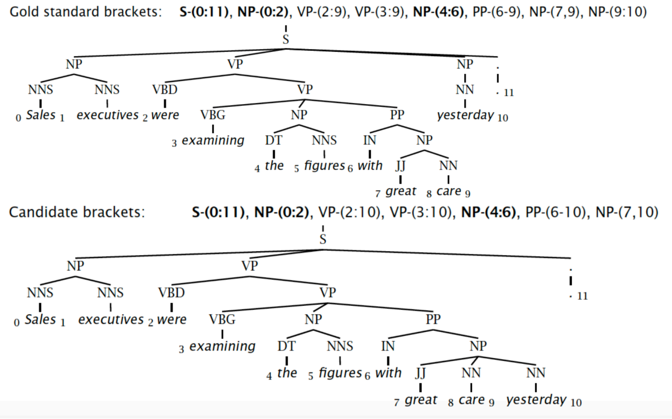
Evaluating Parsers

- How much do the constituents in the hypothesis parse tree look like the constituents in a hand-labeled, reference parse?
- Requires a human-labeled reference (or "gold standard") parse tree for each sentence in the test set
- A constituent in a hypothesis parse of a sentence is labeled correct if there is a constituent in the reference parse with the same starting point, ending point, and non-terminal symbol.

$$\textbf{labeled recall:} = \frac{\text{\# of correct constituents in hypothesis parse of } s}{\text{\# of total constituents in reference parse of } s}$$

$$\textbf{labeled precision:} = \frac{\text{\# of correct constituents in hypothesis parse of } s}{\text{\# of total constituents in hypothesis parse of } s}$$

Evaluating Parsers)



Evaluating Parsers)

Gold standard brackets:

S-(0:11), **NP-(0:2)**, VP-(2:9), VP-(3:9), **NP-(4:6)**, PP-(6-9), NP-(7,9), NP-(9:10)

Candidate brackets:

S-(0:11), **NP-(0:2)**, VP-(2:10), VP-(3:10), **NP-(4:6)**, PP-(6-10), NP-(7,10)

Labeled Precision	3/7 = 42.9%
Labeled Recall	3/8 = 37.5%
LP/LR F1	40.0%
Tagging Accuracy	11/11 = 100.0%

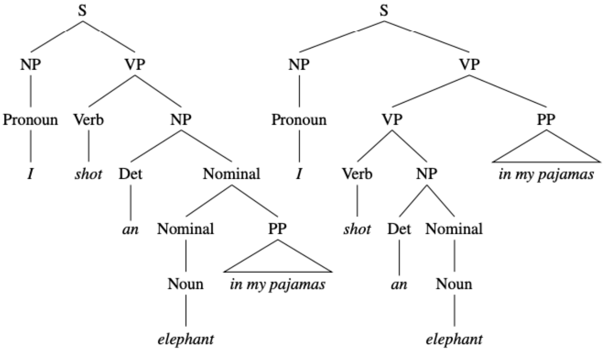
1. Syntactic Parsing
2. Named Entity Recognition
3. Named Entity Recognition
4. Discourse Coherence

How parse trees resolve different kinds of ambiguity

- Ambiguity is the most serious problem faced by syntactic parsers.
- **Structural ambiguity** occurs when the grammar can assign more than one parse to a sentence.
- Groucho Marx: *One morning **I shot an elephant in my pajamas**. How he got into my pajamas I'll never know*

How parse trees resolve different kinds of ambiguity

Ambiguous because the phrase *in my pajamas* can be part of the NP headed by *elephant* or a part of the verb phrase headed by *shot*



Treebanks

- A corpus in which every sentence is annotated with a parse tree is called a treebank
- Treebanks are generally made by running a parser over each sentence and then having the resulting parse hand-corrected by human linguists.
- The most well-known of these is the **Penn Treebank**

```
((S
   (NP-SBJ (DT That)
     (JJ cold) (, ,)
     (JJ empty) (NN sky) )
   (VP (VBD was)
     (ADJP-PRD (JJ full)
       (PP (IN of)
         (NP (NN fire)
           (CC and)
           (NN light) ))))
   (. .) ))
```

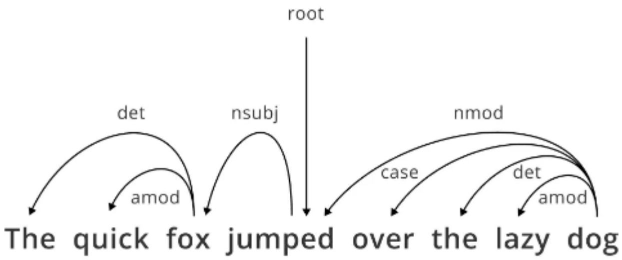
(a)

```
((S
   (NP-SBJ The/DT flight/NN )
   (VP should/MD
     (VP arrive/VB
       (PP-TMP at/IN
         (NP eleven/CD a.m/RB ))
       (NP-TMP tomorrow/NN )))))
```

(b)

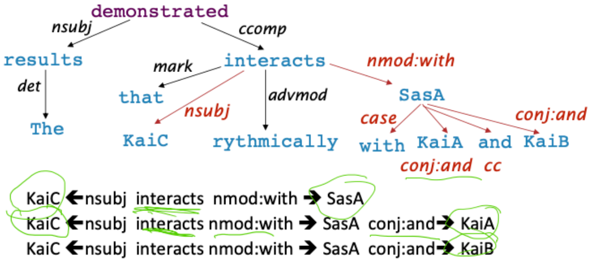
Alternative to constituency structure: dependency structure

- Dependency syntax postulates that syntactic structure consists of relations between lexical items, normally binary asymmetric relations (arrows) called dependencies
- Dependency structure shows which words depend on (modify, attach to, or are arguments of) which other words.



Alternative to constituency structure: dependency structure

- Dependency paths help extract semantic interpretation simple practical example: extracting protein-protein interaction



1. Syntactic Parsing
2. Named Entity Recognition
3. Named Entity Recognition
4. Discourse Coherence

Named Entities

- Named entity, in its core usage, means anything that can be referred to with a proper name. Most commonly used 4 tags:
  - PER (Person): “Marie Curie”
  - LOC (Location): “New York City”
  - ORG (Organization): “Stanford University”
  - GPE (Geo-Political Entity): “Boulder, Colorado”
- Often multi-word phrases
- But the term is also extended to things that aren’t entities: dates, times, prices

Named Entity tagging

The task of named entity recognition (NER):

- find spans of text that constitute proper names
- tag the type of the entity.

NER output

Citing high fuel prices, [$_{ORG}$ **United Airlines**] said [$_{TIME}$ **Friday**] it has increased fares by [$_{MONEY}$ **$6**] per round trip on flights to some cities also served by lower-cost carriers. [$_{ORG}$ **American Airlines**], a unit of [$_{ORG}$ **AMR Corp.**], immediately matched the move, spokesman [$_{PER}$ **Tim Wagner**] said. [$_{ORG}$ **United**], a unit of [$_{ORG}$ **UAL Corp.**], said the increase took effect [$_{TIME}$ **Thursday**] and applies to most routes where it competes against discount carriers, such as [$_{LOC}$ **Chicago**] to [$_{LOC}$ **Dallas**] and [$_{LOC}$ **Denver**] to [$_{LOC}$ **San Francisco**].

Why NER?

- Sentiment analysis: consumers sentiment toward a particular company or person?
- Question Answering: answer questions about an entity?
- Information Extraction: Extracting facts about entities from text.

Why NER is hard

- Segmentation: In POS tagging, no segmentation problem since each word gets one tag.
- In NER we have to find and segment the entities!
- Type ambiguity:

[$_{\text{PER}}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{\text{ORG}}$ Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [$_{\text{LOC}}$ Washington] for what may well be his last state visit.
In June, [$_{\text{GPE}}$ Washington] passed a primary seatbelt law.

BIO tagging scheme

- How can we turn this structured problem into a sequence problem like POS tagging, with one label per word?

[PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago] route.

BIO tagging scheme

[PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago] route.

Words	BIO Label
Jane	B-PER
Villanueva	I-PER
of	O
United	B-ORG
Airlines	I-ORG
Holding	I-ORG
discussed	O
the	O
Chicago	B-LOC
route	O
.	O

BIO tagging scheme

Words	BIO Label
Jane	B-PER
Villanueva	I-PER
of	O
United	B-ORG
Airlines	I-ORG
Holding	I-ORG
discussed	O
the	O
Chicago	B-LOC
route	O
.	O

- **B**: token that begins a span
- **I**: tokens inside a span
- **O**: tokens outside of any span
- # of tags (where n is entity types):
  - 1 O tag,
  - n B tags,
  - n I tags
  - total of 2n+1

BIO Tagging variants: IO and BIOES

[PER Jane Villanueva] of [ORG United] , a unit of [ORG United Airlines Holding] , said the fare applies to the [LOC Chicago] route.

Words	IO Label	BIO Label	BIOES Label
Jane	I-PER	B-PER	B-PER
Villanueva	I-PER	I-PER	E-PER
of	O	O	O
United	I-ORG	B-ORG	B-ORG
Airlines	I-ORG	I-ORG	I-ORG
Holding	I-ORG	I-ORG	E-ORG
discussed	O	O	O
the	O	O	O
Chicago	I-LOC	B-LOC	S-LOC
route	O	O	O
.	O	O	O

Standard algorithms for NER

Supervised Machine Learning given a humanlabeled training set of text annotated with tags

- Hidden Markov Models
- Conditional Random Fields (CRF) / Maximum Entropy Markov Models (MEMM)
- Neural sequence models (RNNs or Transformers) Large Language Models (like BERT), finetuning

1. Syntactic Parsing
2. Named Entity Recognition
3. Named Entity Recognition
4. **Discourse Coherence**

What is Discourse Coherence?

- **Discourse Coherence** refers to the quality of a text or spoken conversation that makes it **semantically unitary** and **meaningful**.
- It is the way in which the sentences and ideas in a text relate to each other to form a whole.
- A coherent text is not just a collection of random sentences; it has a logical flow and structure.
- Coherence is essential for comprehension and interpretation.
- John took a train from Paris to Istanbul. He likes spinach. vs *Jane took a train from Paris to Istanbul. She had to attend a conference*

Types of Coherence

- **Local Coherence**
  - Deals with the relationship between **adjacent sentences** or clauses.
  - Focuses on immediate links and transitions.
  - Often achieved through **cohesion** (lexical and grammatical ties).
- **Global Coherence**
  - Deals with the relationship of all sentences to the **overall topic** or goal of the discourse.
  - Ensures the entire text makes sense as a whole and follows a central theme (e.g., narrative plot, argument structure).

Lexical Cohesion

- Refers to the coherence created by the **choice of words** and the semantic relationships between them.
- Achieved through:
    - **Repetition**: Repeating the same word or phrase.
    - **Synonymy/Antonymy**: Using words with similar or opposite meanings.
    - **Hyponymy/Hypernymy**: Using words that are sub-classes or super-classes (e.g., *dog* is a hyponym of *animal*).
    - **Collocation**: Words that frequently occur together (e.g., *heavy rain*).

Entity-Based Coherence

- Focuses on the **tracking of entities** (people, objects, concepts) throughout the text.
- Achieved primarily through **reference resolution**:
    - **Anaphora**: Using a pronoun or shorter phrase to refer back to an already introduced entity (e.g., *John* arrived. *He* sat down).
    - Coherence is maintained if the reader can easily identify the entities being discussed and how they are referred to in subsequent sentences.

Coherence Relations (Discourse Relations)

- These are the **explicit or implicit links** that hold two segments of text (propositions, clauses, sentences) together.
- They describe **how the content of one segment relates to the content of another**.
- Examples of common relations:
  - **Causality**: (Because, Since)
  - **Contrast**: (However, But)
  - **Elaboration**: (Specifically, For example)
  - **Temporal**: (Then, After)
  - **Condition**: (If, Provided that)

Rhetorical Structure Theory (RST)

- A widely used theory for modeling discourse structure, developed by Mann and Thompson (1988).
- Models the text as a **tree structure** where non-overlapping spans of text are linked by **coherence relations**.
- The basic structure involves:
  - **Nucleus**: The central, essential part of the relation.
  - **Satellite**: The subsidiary, less central part (often provides background, evidence, or context).
- **Goal**: Explain the coherence of a text by identifying its components and the rhetorical relations that hold between them.

Key Concepts in RST

- **Relation:** The link that connects two spans of text.
- **Nucleus (NUC):** The essential, central part of the relation. The speaker's primary intent is often conveyed here.
- **Satellite (SAT):** The subsidiary, less essential part. It supports, elaborates, or provides context for the Nucleus.

1. Reason

- **Definition:** The nucleus is an action carried out by an animate agent and the satellite is the **reason** for the nucleus.
- **Goal:** To explain the motivation for the action in the Nucleus.

`[NUC Jane took a train from Paris to Istanbul.] [SAT She had to attend a conference.]`

- The action (taking the train) is the NUC.
- The motivation (attending a conference) is the SAT.

2. Elaboration

- **Definition:** The satellite gives **additional information or detail** about the situation presented in the nucleus.
- **Goal:** To provide richness, specificity, or context to the Nucleus.

`[NUC Dorothy was from Kansas.] [SAT She lived in the midst of the great Kansas prairies.]`

- The core fact (being from Kansas) is the NUC.
- The detailed description (where she lived) is the SAT.

3. Evidence

- **Definition:** The satellite gives information with the goal of **convincing the reader** to accept the information presented in the nucleus.
- **Note:** Unlike `Reason`, which explains an agent's motivation, `Evidence` supports the **belief** in a proposition.

```
[NUC Kevin must be here.] [SAT His car is parked outside.]
```

- The claim (Kevin is here) is the NUC.
- The supporting observation (car parked outside) is the SAT.

4. Attribution

- **Definition:** The satellite gives the **source of attribution** for an instance of reported speech or thought in the nucleus.
- **Goal:** To credit the source of the reported information.

```
[SAT Analysts estimated] [NUC that sales at U.S. stores declined in the quarter, too]
```

- The reported content (sales declined) is the NUC.
- The source of the report (Analysts estimated) is the SAT.
- **Note:** The SAT precedes the NUC in this example.

5. List

- **Definition:** A **multinuclear** relation where a series of nuclei is given, without contrast or explicit comparison.
- **Note:** In multinuclear relations, **all spans are Nuclei** because they are all central to the writer's purpose.

```
[NUC Billy Bones was the mate; ] [NUC Long John, he was quartermaster]
```

- Both items (Billy Bones' role and Long John's role) are equally important and are merely listed as related facts.

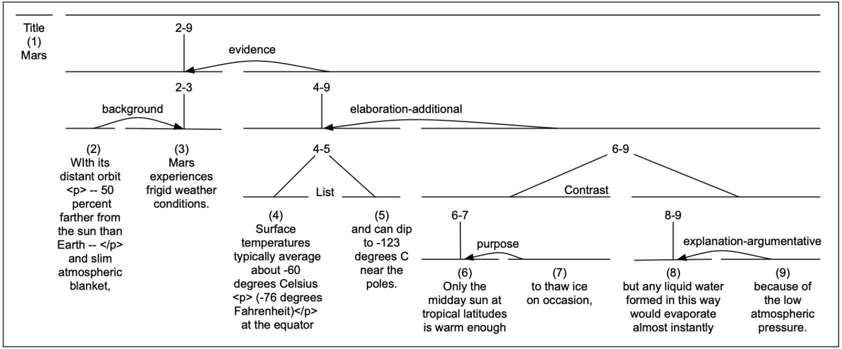
Summary of Relations

Relation	**Function of Satellite**
Reason	Explains the **motive** for an action
Elaboration	Provides **detail** about the NUC
Evidence	Supports the **belief** in the NUC
Attribution	Gives the **source** of the NUC's content
List	Presents **equally weighted items** (Multinuclear)

Example RST parse: Scientific American Article

(24.12) With its distant orbit–50 percent farther from the sun than Earth–and slim atmospheric blanket, Mars experiences frigid weather conditions. Surface temperatures typically average about -60 degrees Celsius (-76 degrees Fahrenheit) at the equator and can dip to -123 degrees C near the poles. Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, but any liquid water formed in this way would evaporate almost instantly because of the low atmospheric pressure.

Example RST parse: Scientific American Article



The Penn Discourse Treebank (PDTB)

- A large-scale corpus annotated with **discourse relations**.
- Distinctive features:
  - Annotation is **local**: Relations are marked only between adjacent arguments.
  - Focuses on **connectives** (e.g., *although, because, and*) and implicit relations.
  - **Arguments** of a relation are entire spans of text (e.g., sentence or clause).
  - **Lexicalized approach**: Relations are anchored to a discourse connective (explicit or implicit).
- PDTB is crucial for training and evaluating automated discourse parsers.

Discourse Parsing

- The computational task of automatically **determining the structure and coherence relations** in a text.
- It is a form of higher-level linguistic analysis, going beyond syntactic parsing.
- **Goal**: To represent the discourse structure explicitly for natural language processing (NLP) applications like:
  - Text Summarization (identifying the central 'Nucleus' parts).
  - Machine Translation (handling cross-sentential dependencies).
  - Question Answering (understanding causal links).
- Parsers often rely on machine learning models trained on corpora like RST-DT (for RST) or the PDTB.

Next class: November 4

Assignment 3: Due November 11

Text generation tasks

- Jurafsky & Martin Chapter 12: Machine Translation
- Jurafsky & Martin Chapter 11: Information Retrieval and Retrieval-Augmented Generation
- A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models
- Automated text simplification: A survey
- Paraphrase Generation: A Survey of the State of the Art
- Commonsense Knowledge Reasoning and Generation with Pre-trained Language Models: A Survey
- Jurafsky & Martin Chapter 15: Chatbots & Dialogue Systems