Carnegie Mellon University
Language
Technologies
Institute

11-411/11-611 Natural Language Processing

Part of Speech Tagging and Named Entity Recognition

David R. Mortensen

February 28, 2023

Language Technologies Institute

Learning Objectives

At the end of this lecture, you should be able to do the following things:

- Define the criteria for classifying parts of speech
- Identify open class and closed class parts of speech
- Define the task of POS tagging (part of speech tagging) and talk about two approaches to it
- Explain how HMMs can be used to tag text for parts of speech
- Be able to implement an HMM decoder for POS tagging using the Viterbi Algorithm
- Define the task of NER (named entity recognition)
- Be able to describe some downstream uses of NER, especially with regard to your projects
- Tag text with BIO NER labels

Two Examples: POS Tagging and NER

Sequences are everywhere in language and many tasks involve classifying the items in those sequences. Two examples:

- Part of Speech Tagging
- Named Entity Recognition

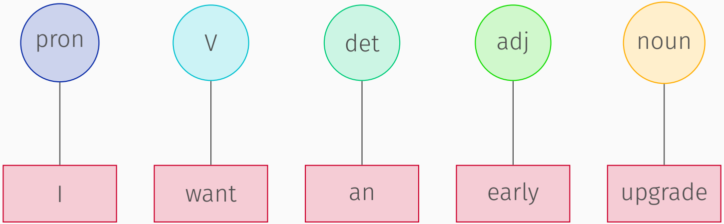
We will use these two tasks as examples of the general sequences labeling task.

Definitions of POS Tagging

Part of speech tagging/labeling:

input A natural language text tokenized into words

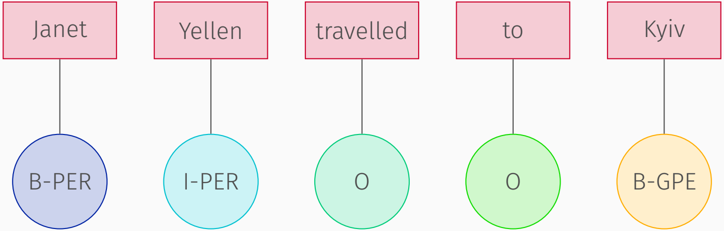
output A sequence of part-of-speech tags, one for each token in the input



Definition of NER

input A natural language text tokenized into words

output A sequence of named entity tags (e.g., BIO tags) with our without type labels, one for each token in the input



Parts of Speech

My cat who lives dangerously no longer has nine lives.

My cat who lives dangerously no longer has nine lives.

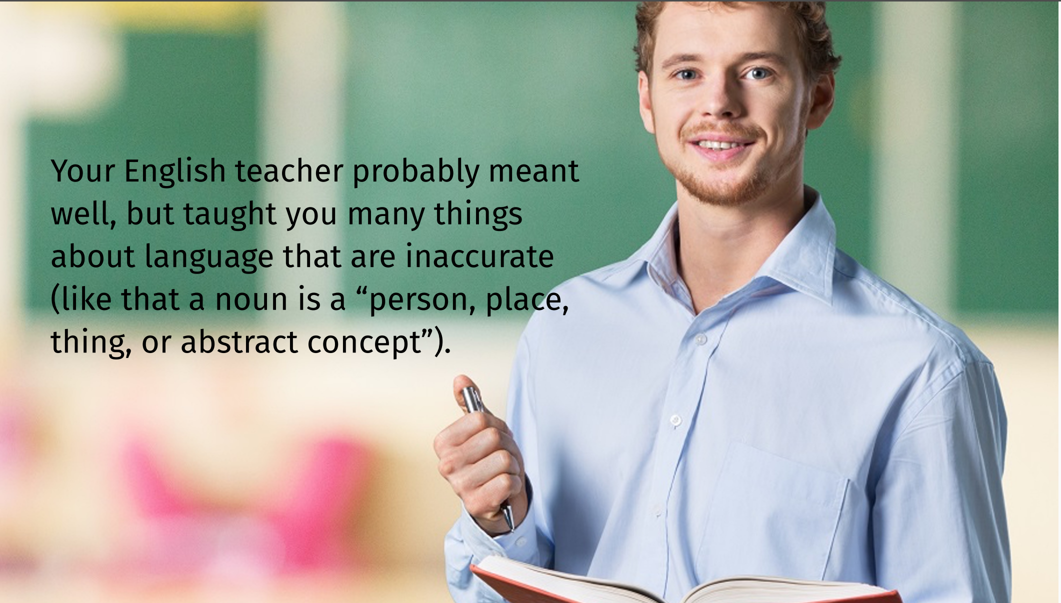
My cat who lives dangerously no longer has nine lives.

lives	/lɪvz/	verb
lives	/lajvz/	noun

Examples of Parts of Speech

PART OF SPEECH	EXAMPLES
noun	dog, cat, professor, exam, fear, loathing, oppression, void, text, Bavarian
verb	enjoy, walk, finish, trust, hug, like, understand, be, text, drink
adjective	nice, happy, red, exciting, ludicrous, funny, ancient, Bavarian
adverb	slowly, quickly, shrewdly, foolishly, boisterously, undercover, yesterday
preposition	to, for, from, under, by
auxiliary verbs	be, have, must, might, will, would
determiner	the, a(n), this, that, my, her
pronouns	he, she, it, this, that
conjunctions	and, but, however, nevertheless, so

Your English Teacher Was a Well-Intentioned Liar



Your English teacher probably meant well, but taught you many things about language that are inaccurate (like that a noun is a “person, place, thing, or abstract concept”).

Criteria for Parts of Speech

Remember the early 20th century American linguists who wanted to document endangered languages? They wanted to define parts of speech in an objective, language-neutral way, so the defined them **distributionally**. This works better than the semantic criteria that your English teacher taught you.

morphology What is the distribution of morphemes within these words?
Same POS $\Rightarrow$ similar morphemes

syntax What is the distribution of words within phrases and sentences?
Same POS $\Rightarrow$ similar roles/contexts

American Structuralists called these "form classes" but we call them "lexical classes" or "grammatical classes" or "parts of speech"

Open Class Parts of Speech

Classes to which neologisms are readily added. In English:

nouns	She trained her neural **models** quickly
verbs	She **trained** her neural models quickly
adjectives	She trained her **neural** models quickly
adverbs	She trained her neural models **quickly**

Closed Class Parts of Speech

Classes to which neologisms are not readily added. In English:

prepositions	**After** two minutes, he had taken off his glasses and had tossed them **through** the window
determiners	After two minutes, he had taken off his glasses and had tossed them through **the** window
conjunction	After two minutes, he had taken off his glasses **and** had tossed them through the window
auxiliary verbs	After two minutes, he **had** taken off his glasses and **had** tossed them through the window
particles	After two minutes, he had taken **off** his glasses and had tossed them through the window
numerals	After **two** minutes, he had taken off his glasses and had tossed them through the window

Open Class Parts of Speech Defined

Classes to which neologisms are readily added. In English:

nouns	can be both subjects and objects of verbs and objects of prepositions, (usually) be singular or plural, have determiners, be modified by adjectives, and be possessed
verbs	can take noun phrases as arguments and tense morphology and can be modified by adverbs
adjectives	can modify nouns and take comparative and superlative morphology where allowed by prosody
adverbs	can modify verbs, adjectives, or other adverbs

Closed Class Parts of Speech Defined

Classes to which neologisms are not readily added. In English:

prepositions	occur before noun phrases, connecting them syntactically to larger phrases
determiners	occur at the beginning of noun phrases
conjunction	join phrases, clauses, and sentences
auxiliary verbs	occur before (non-finite) main verbs
particles	are associated with a verb and are "moveable" (e.g. *He tore* ***off*** *his shirt* versus *He tore his shirt* ***off***
numerals	are distributed in some ways like nouns and in others like adjectives

What about Pronouns?

Pronouns are generally considered, in English, to be a closed class—it is not easy to add new items to it.

What are we to make of **neopronouns** like *xe* and *xem* or *ze* and *hir*?

point Their existence suggests that pronouns are not a closed class

counterpoint The difficulty with which people learn or use them suggests that pronouns are a closed class

In some languges (e.g., Thai) pronouns clearly *are* an open class.

Part of Speech Tagging

The POS Labeling Task

input A natural language text tokenized into words

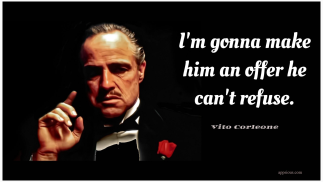
output A sequence of part-of-speech tags, one for each token in the input

What can we accomplish with POS tagging?

POS Tagging is a Disambiguation Task

Consider the following sentences:

I	’m	gonna	make	him	an	offer	he	can	’t	refuse
PRO	V	AUX	V	PRO	DET	N	PRO	AUX	ADV	V
			N			V				N



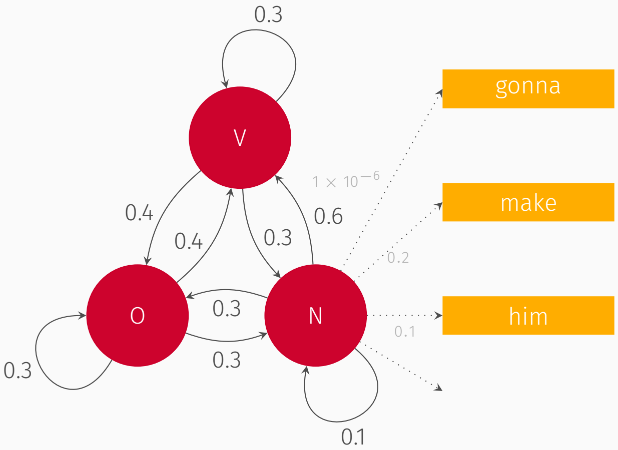
There are eight different ways of tagging this sentence if words are taken out of context. POS Tagging task: **choose the best of these**.

Tagging Parts of Speech with HMMs

Simplifying Assumptions

For sake of interpretability, let us concern ourselves with only three parts of speech: NOUN+PRONOUN (N), VERB+AUXILIARY VERB (V) and OTHER (O). Assume that we have an HMM $\lambda = (A, B)$ as described in the following two slides.

An HMM for POS



Transition Probabilities

$$A = \begin{array}{c|ccc} & \text{N} & \text{V} & \text{O} \\ \hline \text{N} & 0.1 & 0.6 & 0.3 \\ \text{V} & 0.3 & 0.3 & 0.4 \\ \text{O} & 0.3 & 0.4 & 0.3 \end{array}$$

Emission Probabilities and Initial Probabilities

$B =$

	I	m	gonna	make	him	an	offer	he	can	t	refuse
N	0.1	0.00001	0.00001	0.2	0.1	0.00001	0.2	0.1	0.1	0.00001	0.19996
V	0.00001	0.1	0.2	0.2	0.00001	0.00001	0.05	0.00001	0.19995	0.00001	0.25
O	0.00001	0.00001	0.00001	0.00001	0.00001	0.5	0.00001	0.00001	0.00001	0.49991	0.00001

$$\pi = [0.5, 0.2, 0.3]$$

Reminder: Viterbi Decoding

function VITERBI(observations $O = o_1, o_2, \ldots, 0_T$, state-graph of length N)

  $V[N, T] \leftarrow$ empty path probability matrix

  $B[N, T] \leftarrow$ empty backpointer matrix

  for each $s \in 1..N$ **do**

    $V[s, 1] \leftarrow \pi_s \cdot b_s(o_1)$

    $B[s, 1] \leftarrow 0$

  for each $t \in 2..T$ **do**

    for each $s \in 1..N$ **do**

      $V[s, t] \leftarrow \max_{s'=1}^{N} V[s', t-1] \cdot a_{s',s} \cdot b_s(o_t)$

      $B[s, t] \leftarrow \operatorname{argmax}_{s'=1}^{N} V[s', t-1] \cdot a_{s',s} \cdot b_s(o_t)$

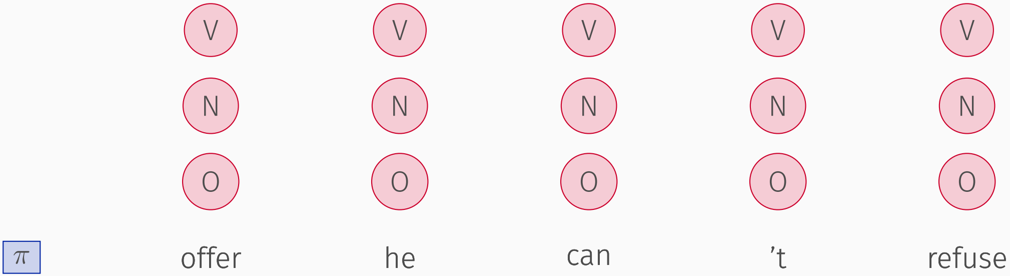
  $bestpathprob \leftarrow \max_{s=1}^{N} V[s, T]$

  $bestpathpointer \leftarrow \operatorname{argmax}_{s=1}^{N} V[s, T]$

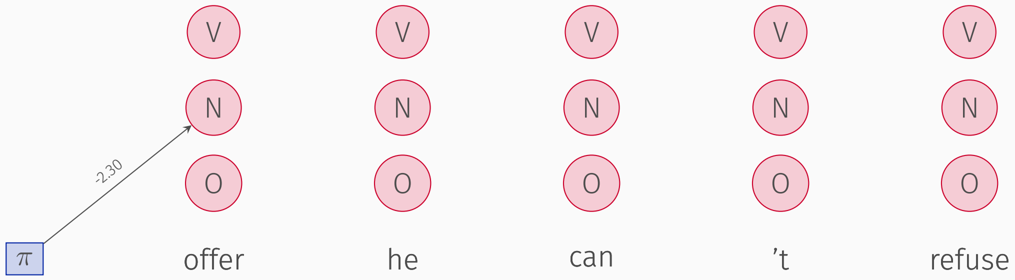
  $bestpath \leftarrow$ path starting at *bestpathpointer* that follows b to states back in time.

  return *bestpath*, *bestpathprob*

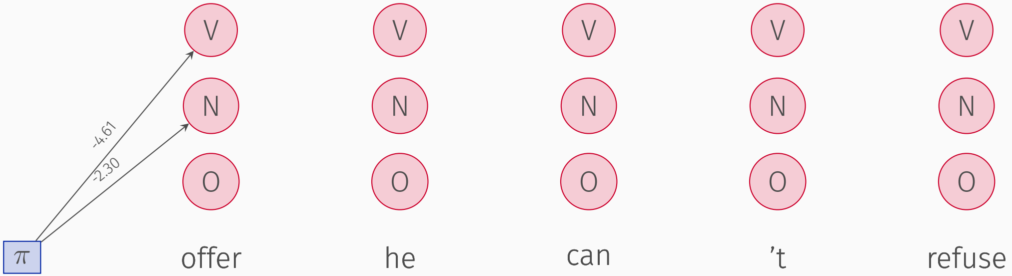
A Trellis You Can't Refuse



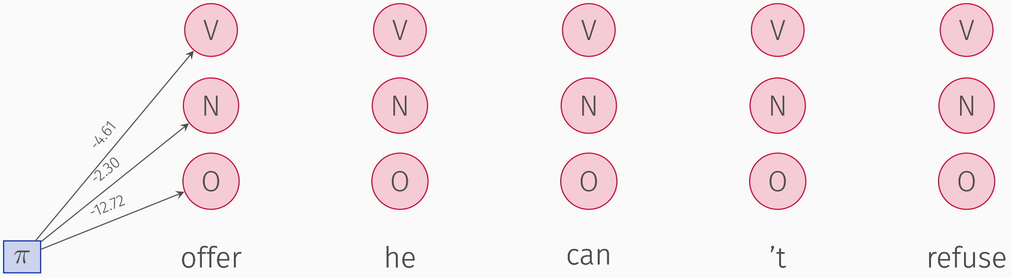
A Trellis You Can't Refuse



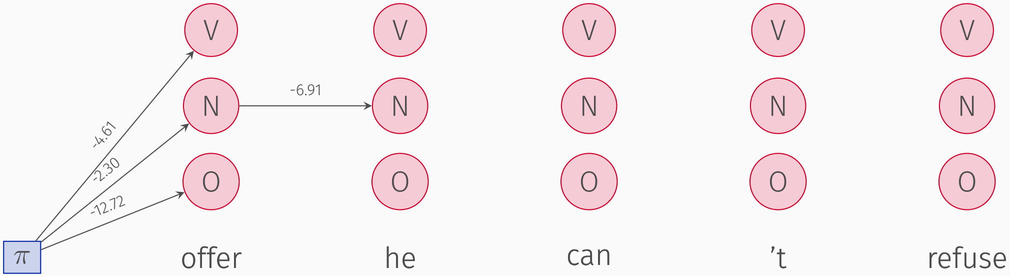
A Trellis You Can't Refuse



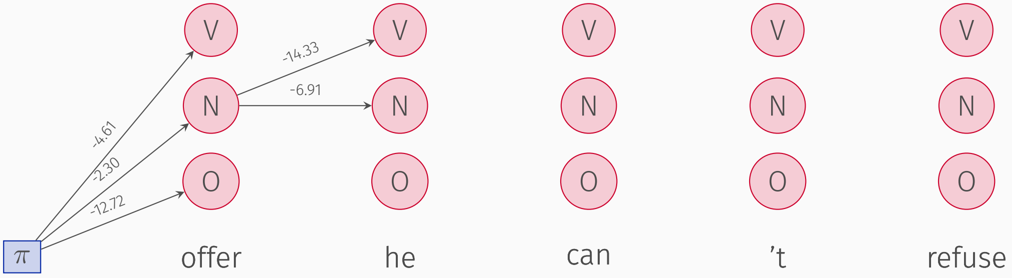
A Trellis You Can't Refuse



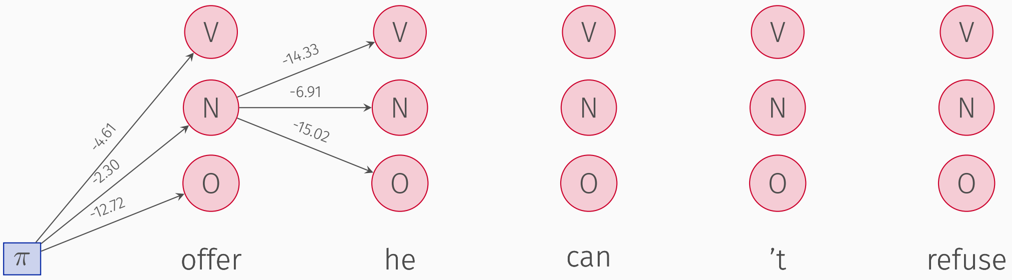
A Trellis You Can’t Refuse



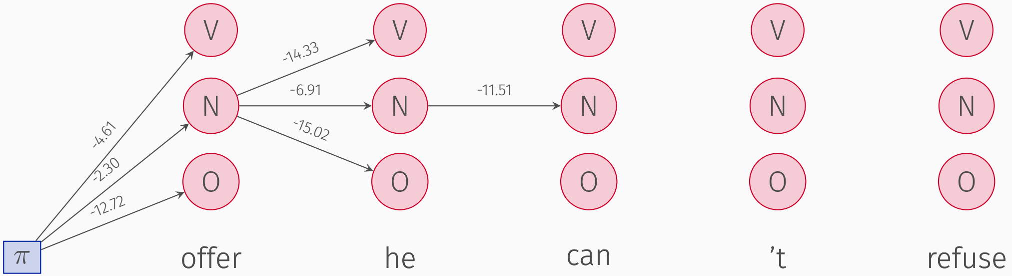
A Trellis You Can't Refuse



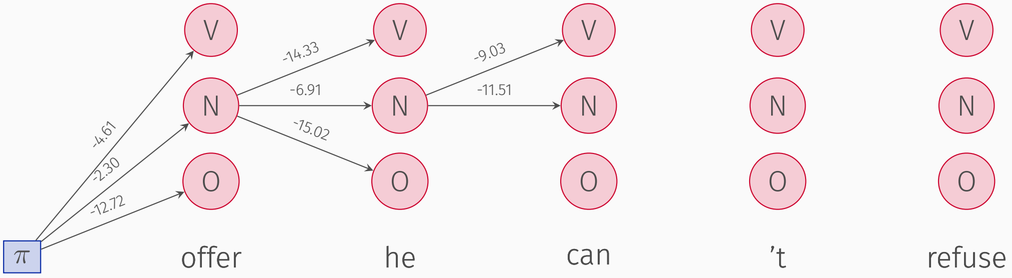
A Trellis You Can’t Refuse



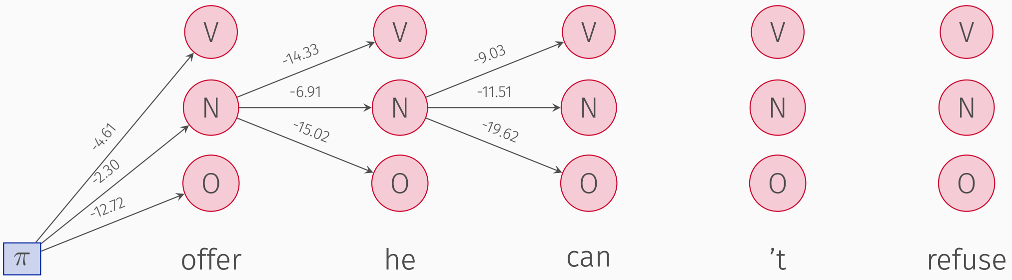
A Trellis You Can’t Refuse



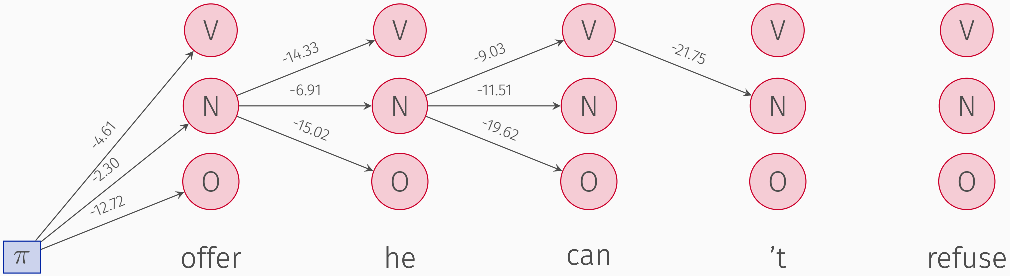
A Trellis You Can't Refuse



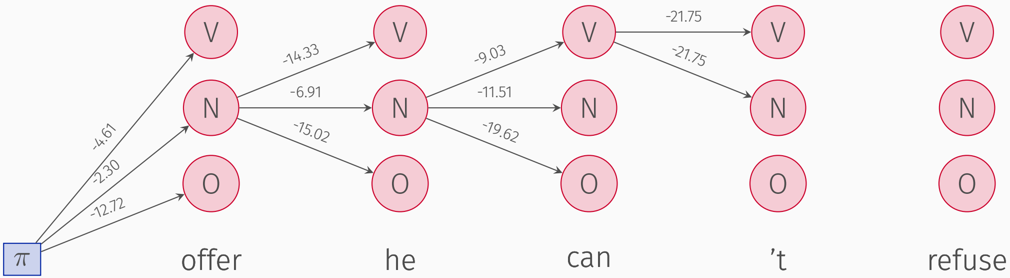
A Trellis You Can't Refuse



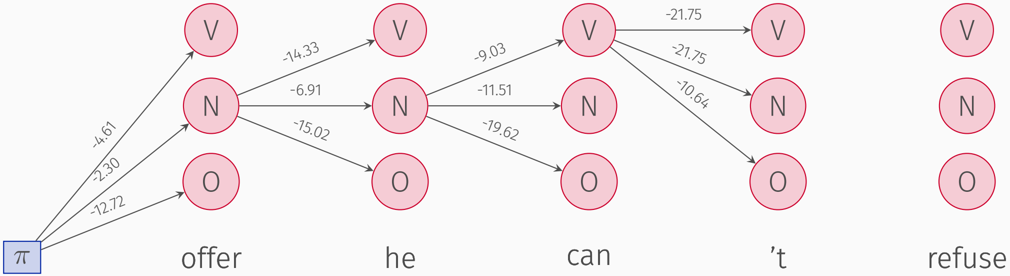
A Trellis You Can't Refuse



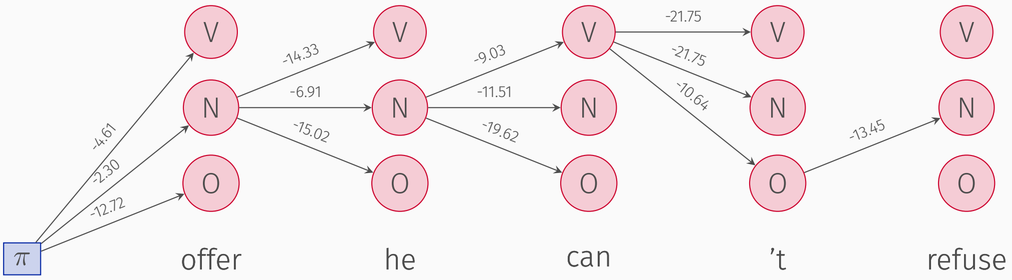
A Trellis You Can't Refuse



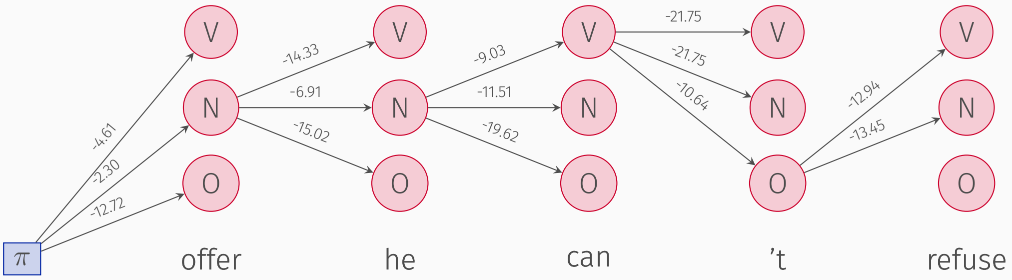
A Trellis You Can't Refuse



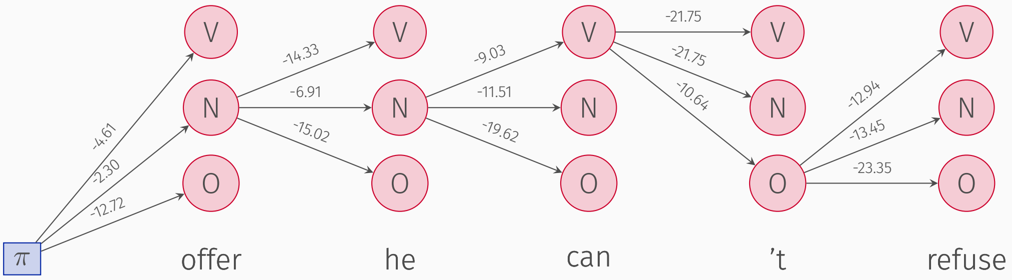
A Trellis You Can't Refuse



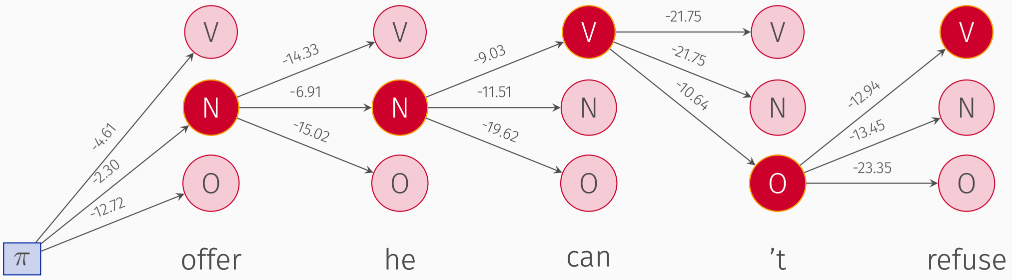
A Trellis You Can't Refuse



A Trellis You Can't Refuse



A Trellis You Can't Refuse



Conditional Random Fields

Conditional Random Fields are Bidirectional

- Conditional random fields are like HMMs in that all information is local
- However, CRFs look ahead rather than behind
- The optimal path is computed dynamically, but not in the same way as for HMMs
- Since POS (and NER) information often comes from both directions, CRFs are valuable
- Sometimes they are used in conjunction with CNNs and bidirectional LSTMs to build powerful sequence labeling models (e.g., for NER).

Named Entity Recognition

Named Entity Recognition is Identifying Name Spans in Text

input A natural language text tokenized into words

output A sequence of named entity tages (e.g., BIO tags), with our without type labels, one for each token in the input

The Task of Named Entity Recognition

Elizabeth Warren, the liberal firebrand who emerged as a top Democratic contender for the White House on the strength of an anti-corruption platform backed by a dizzying array of policy proposals, ended her campaign on Thursday. A former bankruptcy law professor who forged a national reputation as a scourge of Wall Street even before entering politics, Warren had banked on a strong showing on Super Tuesday after a string of disappointing finishes in the early states. But she trailed far behind front-runners Bernie Sanders and Joe Biden, placing third in her home state of Massachusetts, which she continues to represent in the U.S. Senate.

Label certain kinds of proper nouns:

- Personal names
- Organizations
- Geopolitical entities
- Locations
- Dates
- Named natural phenomena (e.g., hurricanes)
- etc.

Encoding NER with BIO

President Donald Trump met with local leaders and federal responders shortly after landing at an **Air Force** base in **Carolina**, **Puerto Rico**, for what was supposed to be a briefing on the situation on the island. Instead, **Trump** turned it into an opportunity to congratulate himself and the federal government's response to the disaster.... He downplayed throughout his remarks how dire things are in **Puerto Rico**, where more than half of the people don't have power, running water, or cellphone service two weeks after **Hurricane Maria**, a Category 4 storm, tore through the island.

- **B:** beginning of NE
- **I:** inside of NE
- **O:** outside of NE

...	...
B-PER	President
I-PER	Donald
I-PER	Trump
O	met
O	with
O	local
O	leaders
O	and
O	federal
O	responders
O	shortly
O	after
O	landing
O	at
O	an
B-ORG	Air
I-ORG	Force
O	base
O	in
B-LOC	Carolina
O	,
B-LOC	Puerto
I-LOC	Rico
...	...

Some Named Entity Types

Different annotation schemes for NER use different types. Common types include:

- PER—person
- ORG—Organization
- LOC—Location
- GPE—Geopolitical Entity
- FAC—Facility
- NAT—Natural phenomenon

In biomedical NER, named entities may include:

- Prescription medications
- Proteins
- Genes
- Diseases
- Cell-type
- Cell-line

These are only tagged when they are proper names

NER Features

Various kinds of features have been used for NER:

- Orthographic cues like capitalization
- Presence of a word in a gazetteer (collection of names)
- Part of speech
- Preceding and following words (e.g., some prepositions often occur before LOCs and GPEs)
- Titles (e.g., “Dr.” usually occurs at the beginning of names
- Morphology
- Embeddings

NER is a Fundamental Information Extraction Task

Many other tasks rely upon NER:

- Entity linking
- Relation extraction
- Coreference resolution
- Question answering
- etc.,

Whenever you need to find all the names, and know what kind of names they are, NER should be your first tool.

NER and RNNs

As mentioned above, NER systems sometimes employ CNNs, Bi-LSTMs, and CRFs

- CNNs to do local feature extraction
- Bi-LSTMs to encode long-distance relationships in either direction
- CRFs to do the high-level sequence modeling

Evaluation: Micro- and Macro- Averaged Precision, Recall, and F1

Because NER is a multiclass classification task (like POS tagging), it is evaluated using micro- or macro-averaged precision and recall

- **micro-averaged** first sum true positives, false positives, and false negatives, then compute precision and recall as usual
- **macro-averaged** compute the precision and recall for each class, then divide by the number of classes

When to Use which Kind of Average Scoring?

When to use which?

- **micro-averaged** scores treat every observation equally. Use when frequency of categories is BALANCED.
- **macro-averaged** scores treat every category equally. Use when frequency of categories is IMBALANCED.

Consider an Example with Seven Classes

Assume the NER types PER, ORG, and DATE ($\times$ B, I) plus O:

	Reference	Hypothesis
Congress	B-ORG	B-ORG
passed	O	O
President	O	B-PER
Joe	B-PER	I-PER
Biden's	I-PER	I-PER
spending	O	O
package	O	O
on	O	O
Tuesday	B-DATE	B-DATE

Consider an Example with Nine Classes

	TP	FP	FN
O	4	1	1
B-PER	1	1	1
I-PER	0	1	0
B-ORG	1	0	0
B-DATE	1	0	0

	Reference	Hypothesis
Congress	B-ORG	B-ORG
passed	O	O
President	O	B-PER
Joe	B-PER	I-PER
Biden's	I-PER	I-PER
spending	O	O
package	O	O
on	O	O
Tuesday	B-DATE	B-DATE

Computing Micro-Averaged Precision and Recall

	TP	FP	FN
O	4	1	1
B-PER	1	1	1
I-PER	0	1	0
B-ORG	1	0	0
B-DATE	1	0	0

$$\begin{aligned}\text{Precision}_{\text{micro}} &= \frac{\sum_i^n TP_i}{\sum_i^n TP_i + \sum_i^n FP_i} \\ &= \frac{4+1+1+1}{(4+1+1+1)+(1+1+1)} \\ &= \frac{6}{10} = 0.60\end{aligned}$$

$$\begin{aligned}\text{Recall}_{\text{micro}} &= \frac{\sum_i^n TP_i}{\sum_i^n TP_i + \sum_i^n FP_i} \\ &= \frac{4+1+1+1}{(4+1+1+1)+(1+1)} \\ &= \frac{6}{9} = 0.67\end{aligned}$$

Micro-Averaged Metrics

In summary:

$$\text{Precision}_{\text{micro}} = \frac{\sum_i^n TP_i}{\sum_i^n TP_i + \sum_i^n FP_i} \tag{1}$$

$$\text{Recall}_{\text{micro}} = \frac{\sum_i^n TP_i}{\sum_i^n TP_i + \sum_i^n FN_i} \tag{2}$$

Computing Macro-Averaged Precision and Recall

	Precision	Recall
O	$\frac{4}{5} = 0.8$	$\frac{4}{5} = 0.8$
B-PER	$\frac{1}{2} = 0.5$	$\frac{1}{2} = 0.5$
I-PER	$\frac{0}{1} = 0.0$	$\frac{0}{0} = 1.0$
B-ORG	$\frac{1}{1} = 1.0$	$\frac{1}{1} = 1.0$
B-DATE	$\frac{1}{1} = 1.0$	$\frac{1}{1} = 1.0$
-		

$$\begin{aligned} \text{Precision}_{\text{macro}} &= \frac{\sum_i^n \text{Precision}_i}{n} \\ &= \frac{0.8 + 0.5 + 0.0 + 1.0 + 1.0}{7} \\ &= 0.47 \end{aligned}$$

$$\begin{aligned} \text{Recall}_{\text{macro}} &= \frac{\sum_i^n \text{Recall}_i}{n} \\ &= \frac{0.8 + 0.5 + 1.0 + 1.0 + 1.0}{7} \\ &= 0.61 \end{aligned}$$

Macro-Averaged Metrics

In summary:

$$\text{Precision}_{\text{macro}} = \frac{\sum_i^n \text{Precision}_i}{n} \tag{3}$$

$$\text{Recall}_{\text{macro}} = \frac{\sum_i^n \text{Recall}_i}{n} \tag{4}$$

NER and Your Project

For many approaches to QG and QA, NER is important:

- Tagging sentences for transformations into WH-questions (who, what, when, where, ...)
- Finding names corresponding to WH-words
- Other things, too.

There are a variety of good NER taggers for English:

- SpaCy
- Stanza
- etc.,

Questions?