

CHAPTER 8: Sequence Labeling for Parts of Speech and Named Entities

Instructor: PhD. Nguyen Thi Quy

Group 5: 8.3 - 8.4.3

November 22, 2023

① 8.3 Named Entities and Named Entity Tagging

② 8.4 HMM Part-of-Speech Tagging

8.4.1 Markov Chains

8.4.2 The Hidden Markov Model

8.4.3 The components of an HMM tagger

① 8.3 Named Entities and Named Entity Tagging

② 8.4 HMM Part-of-Speech Tagging

8.4.1 Markov Chains

8.4.2 The Hidden Markov Model

8.4.3 The components of an HMM tagger

Named entity, in its core usage, means anything that can be referred to with a proper name. Most common 4 tags:

- PER (Person): “Marie Curie”
- LOC (Location): “New York City”
- ORG (Organization): “Stanford University”
- GPE (Geo-Political Entity): “Boulder, Colorado”

Named Entities

- 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.

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: consumer's 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

[PER Washington] was born into slavery on the farm of James Burroughs.

[ORG Washington] went up 2 games to 1 in the four-game series.

Blair arrived in [LOC Washington] for what may well be his last state visit.

In June, [GPE Washington] passed a primary seatbelt law.

BIO Tagging

[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

- * 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 $2n+1$

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 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

Supervised Machine Learning given a human-labeled 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), finetuned

① 8.3 Named Entities and Named Entity Tagging

② 8.4 HMM Part-of-Speech Tagging

8.4.1 Markov Chains

8.4.2 The Hidden Markov Model

8.4.3 The components of an HMM tagger

Introducing HMM

The Hidden Markov Model - HMM

- A statistical model with unknown parameters that must be determined from known parameters.
- Extends from the mathematical model: **Markov Chains**.

Applications

- Sequence labeling: NER, POS tagging
- Optical Character Recognition (OCR)
- Speech recognition
- Bioinformatics

Markov chains

Markov chains

A model that tells us something about the probabilities of sequences of random variables, states

- Sequence of states with a temporal order
- States can take values from any discrete set of values.
- **Markov assumption:** When predicting the future, the past doesn't matter

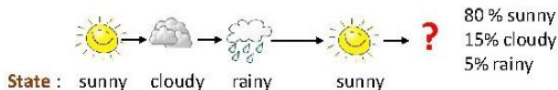
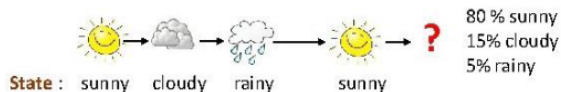


Figure: AA. Markov

Markov assumption

When predicting the future, the past doesn't matter, only the present

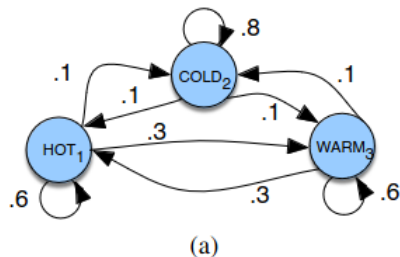


Markov assumption: $P(q_i = a | q_1 \dots q_{i-1}) = P(q_i = a | q_{i-1})$

Components of the Markov chains

- $Q = q_1 q_2 \dots q_n$: a set of N **states**
- $A = a_{11} a_{12} \dots a_{N1} \dots a_{NN}$: a **transition probability matrix** A , each a_{ij} representing the probability of moving from state i to state j
- $\pi = \pi_1, \pi_2, \dots, \pi_n$: an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i .

Markov chains



Markov Chain in Figure (a)

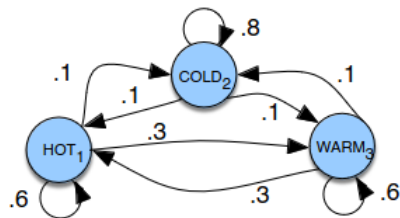
- $Q = \{HOT, COLD, WARM\}$
- Transition probability matrix A :

	π	HOT	COLD	WARM
HOT	0.1	0.6	0.1	0.3
COLD	0.7	0.1	0.8	0.1
WARM	0.2	0.3	0.1	0.6

- Initial probability distribution

$$\pi = [0.1, 0.7, 0.2]$$

Markov chains



(a)

Calculate the probabilities of

- 1 hot hot hot hot
- 2 cold hot cold hot

Markov Chain in Figure (a)

- $Q = \{HOT, COLD, WARM\}$
- Transition probability matrix A :

	π	HOT	COLD	WARM
HOT	0.1	0.6	0.1	0.3
COLD	0.7	0.1	0.8	0.1
WARM	0.2	0.3	0.1	0.6

- Initial probability distribution

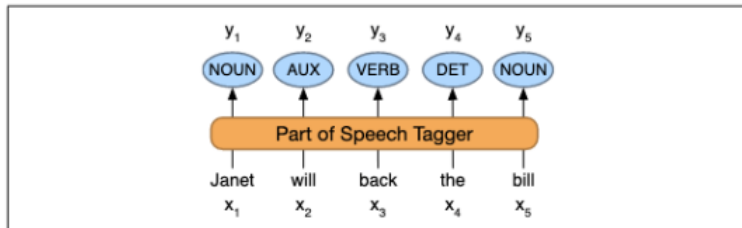
$$\pi = [0.1, 0.7, 0.2]$$

The Hidden Markov Model

A hidden Markov model (HMM) allows us to talk about both observed events and hidden events.

Unobservable Events:

- Part-of-speech
- Entity type



The Hidden Markov Model

$Q = q_1 q_2 \dots q_N$	a set of N states
$A = a_{11} \dots a_{ij} \dots a_{NN}$	a transition probability matrix A , each a_{ij} representing the probability of moving from state i to state j , s.t. $\sum_{j=1}^N a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \dots o_T$	a sequence of T observations , each one drawn from a vocabulary $V = v_1, v_2, \dots, v_V$
$B = b_i(o_t)$	a sequence of observation likelihoods , also called emission probabilities , each expressing the probability of an observation o_t being generated from a state q_i
$\pi = \pi_1, \pi_2, \dots, \pi_N$	an initial probability distribution over states. π_i is the probability that the Markov chain will start in state i . Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^n \pi_i = 1$

Figure: Components of Hidden Markov Model

First order Hidden Markov Model

A first-order HMM instantiates two simplifying assumptions

- 1 The probability of a particular state depends only on the previous state

Markov Assumption: $P(q_i | q_1, \dots, q_{i-1}) = P(q_i | q_{i-1})$

- 2 The probability of an output observation depends only on the state that produced it and not on any other states.

Independence: $P(o_i | q_1, \dots, q_i, \dots, q_T, o_1, \dots, o_i, \dots, o_T) = P(o_i | q_i)$

The Hidden Markov Model

A sample HMM for the ice cream task.

- The two hidden states (H and C) correspond to **hot** and **cold** weather,
- The observations $O = 1, 2, 3$: number of ice creams eaten by Jason on a given day

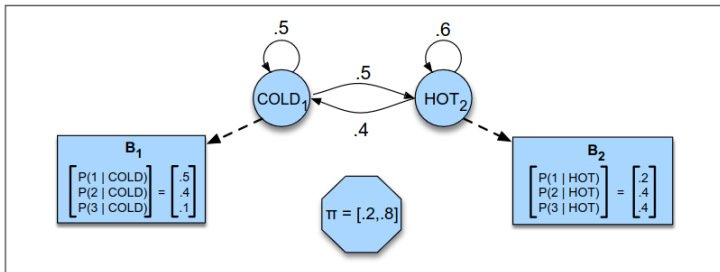


Figure: A hidden Markov model for relating numbers of ice creams eaten by Jason (the observations) to the weather (H or C, the hidden variables).

A model in Natural Language Processing based on HMM, used for labeling elements in a sequence.

HMM Tagger consists of 2 components:

- ① A: The probability of a tag occurring given the previous tag
- ② B: The probability, given a tag, that it will be associated with a given word

The probability of a tag occurring given the previous tag

$$P(t_i | t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}$$

Example - In the WSJ corpus:

- MD occurs **13124** times
- MD is followed by VB **10471** times

Tag transition probability MD - VB:

$$P(VB | MD) = \frac{C(MD, VB)}{C(MD)} = \frac{10471}{13124} = 0.8$$

The probability of a word occurring associated with a tag

$$P(w_i|t_i) = \frac{C(t_i, w_i)}{C(t_i)}$$

Example - In the WSJ corpus:

- MD occurs **13124** times
- MD is associated with *will* **4046** times

Tag transition probability MD - VB:

$$P(\text{will}|MD) = \frac{C(MD, \text{will})}{C(MD)} = \frac{4046}{13124} = 0.31$$



Speech and Language Processing (3rd ed. draft)

Dan Jurafsky and James H. Martin

Part I: Fundamental Algorithms, *Chapter 8: Sequence Labeling for Parts of Speech and Named Entities*

Thanks for listening!

Q&A section