School of CIT
Social Computing
Research Group

Technical
University
of Munich

# Natural Language Processing
# IN2361

Prof. Dr. Georg Groh

# Chapter 8
# Sequence Labeling for Parts of Speech and Named Entities

- content is based on [1]
- certain elements (e.g. equations or tables) were taken over or taken over in a modified form from [1]
- citations of [1] or from [1] are omitted for legibility
- errors are fully in the responsibility of Georg Groh
- BIG thanks to Dan and James for a great book!

# Part-of-Speech-Tags

- Parts-of-Speech (POS, word classes, syntactic categories)
- Examples: noun, pronoun, verb, adjective, ….
- important for
  - language models ("nouns are preceded by determines or adjectives"),
  - information extraction tasks such as Named Entity Recognition and Classification,
  - stemming,
  - auto-summarization,
  - pronunciation (e.g. CONtent vs conTENT)
  - etc.

# Part-of-Speech-Tags

- POS: based on not primarily semantic categories (adjective ←→ property of smth) but rather
  - syntactic categories / functions (e.g. distributional properties (which other words usually in neighborhood)) and
  - morphological categories/functions (e.g. to carry similar suffixes)

- closed class (function words (e.g. *of, it*); fixed members (e.g. prepositions)) vs.
  open class (nouns, verbs, adjectives, adverbs; e.g. new nouns are continually created)

# POS Overview

| | Tag | Description | Example |
|---|---|---|---|
| Open Class | **ADJ** | Adjective: noun modifiers describing properties | *red*, *young*, *awesome* |
| | **ADV** | Adverb: verb modifiers of time, place, manner | *very*, *slowly*, *home*, *yesterday* |
| | **NOUN** | words for persons, places, things, etc. | *algorithm*, *cat*, *mango*, *beauty* |
| | **VERB** | words for actions and processes | *draw*, *provide*, *go* |
| | **PROPN** | Proper noun: name of a person, organization, place, etc.. | *Regina*, *IBM*, *Colorado* |
| | **INTJ** | Interjection: exclamation, greeting, yes/no response, etc. | *oh*, *um*, *yes*, *hello* |
| Closed Class Words | **ADP** | Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation | *in, on, by, under* |
| | **AUX** | Auxiliary: helping verb marking tense, aspect, mood, etc., | *can, may, should, are* |
| | **CCONJ** | Coordinating Conjunction: joins two phrases/clauses | *and*, *or*, *but* |
| | **DET** | Determiner: marks noun phrase properties | *a, an, the, this* |
| | **NUM** | Numeral | *one, two, first, second* |
| | **PART** | Particle: a preposition-like form used together with a verb | *up, down, on, off, in, out, at, by* |
| | **PRON** | Pronoun: a shorthand for referring to an entity or event | *she, who, I, others* |
| | **SCONJ** | Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement | *that*, *which* |
| Other | **PUNCT** | Punctuation | ; , () |
| | **SYM** | Symbols like $ or emoji | $, % |
| | **X** | Other | asdf, qwfg |

**Figure 8.1** The 17 parts of speech in the Universal Dependencies tagset (Nivre et al., 2016a). Features can be added to make finer-grained distinctions (with properties like number, case, definiteness, and so on).

# Part-of-Speech-Tags

- Nouns:
  - occur with determiners (*a goat, its bandwidth*)
  - can take possessives (*husband's house*)
  - may occur in plural (*goats, hounds*)
  - Proper Nouns: specific entities, no *the* (*Regina, IBM, Colorado*) (usually capitalized)
  - Common Nouns:
    - Count Nouns: *one goat, two goats*
    - Mass Nouns: *snow, salt, communism*

- Verbs:
  - ←→ actions, processes, smth. dynamic,…
  - may be inflected: *eat, eats, eating, eaten*

- Adjectives
  - ←→ properties, qualities,…
  - *beautiful, tall, small*

# Part-of-Speech-Tags

- Adverbs:
  - modify something: ***Unfortunately***, *John walked* ***home extremely slowly yesterday***
  - directional adverbs / locative adverbs: *home, here, downhill*
  - degree adverbs: *extremely, very, somewhat*
  - manner adverbs: *slowly, slinkily, delicately*
  - temporal adverbs: *yesterday, Monday*

- Prepositions:
  - occur before noun phrases: ***by*** *the house,* ***on*** *time,* ***with*** *gusto,* ***at*** *the gate*
  - indicate spatial, or temporal, or other relations

- Particle:
  - occur with verbs: *hand the paper* ***over****, throw the ball* ***at***
  - together with verb: phrasal verb (with non-compositional meaning): *turn down == reject, rule out == eliminate, go on == continue*

# Part-of-Speech-Tags

- Determiners:
  - especially articles: definite: *the*; indefinite: *a, an*
  - also*: this, that, …*

- Conjunctions:
  - join phrases, sentences, clauses
  - Coordinating conjunctions*: and, or*
  - Subordinating conjunctions (Complementizers)*: I thought* ***that*** *you might fail*

- Pronouns:
  - shorthand referring to noun phrase etc.
  - Personal pronoun: *you, I, he, she, it*
  - Possessive pronoun: *your, mine, his, her, its, one's*
  - Wh-pronouns: *what, whom, whoever, why*

# Part-of-Speech-Tags

- Auxiliary verbs:
  - mark semantic features of verbs: *can, do, may, should, are, have*: whether action is completed, negated, necessary, possible, suggested, desired, ….
  - Copula *be* : connects: *he **is** a duck*
  - Modal verbs: *can, must*

- Other classes:
  - Interjections *oh, hey, um, hmmm*
  - Negatives *no, not*
  - Politeness markers *please, thank you*
  - Greetings *hello, goodbye*
  - …

# Penn Treebank POS Tags

| Tag | Description | Example | Tag | Description | Example | Tag | Description | Example |
|---|---|---|---|---|---|---|---|---|
| CC | coord. conj. | *and, but, or* | NNP | proper noun, sing. | *IBM* | TO | "to" | *to* |
| CD | cardinal number | *one, two* | NNPS | proper noun, plu. | *Carolinas* | UH | interjection | *ah, oops* |
| DT | determiner | *a, the* | NNS | noun, plural | *llamas* | VB | verb base | *eat* |
| EX | existential 'there' | *there* | PDT | predeterminer | *all, both* | VBD | verb past tense | *ate* |
| FW | foreign word | *mea culpa* | POS | possessive ending | *'s* | VBG | verb gerund | *eating* |
| IN | preposition/ subordin-conj | *of, in, by* | PRP | personal pronoun | *I, you, he* | VBN | verb past partici- ple | *eaten* |
| JJ | adjective | *yellow* | PRP$ | possess. pronoun | *your, one's* | VBP | verb non-3sg-pr | *eat* |
| JJR | comparative adj | *bigger* | RB | adverb | *quickly* | VBZ | verb 3sg pres | *eats* |
| JJS | superlative adj | *wildest* | RBR | comparative adv | *faster* | WDT | wh-determ. | *which, that* |
| LS | list item marker | *1, 2, One* | RBS | superlatv. adv | *fastest* | WP | wh-pronoun | *what, who* |
| MD | modal | *can, should* | RP | particle | *up, off* | WP$ | wh-possess. | *whose* |
| NN | sing or mass noun | *llama* | SYM | symbol | *+,%, &* | WRB | wh-adverb | *how, where* |

There/PRO/EX are/VERB/VBP 70/NUM/CD children/NOUN/NNS
there/ADV/RB ./PUNC/.

Preliminary/ADJ/JJ findings/NOUN/NNS were/AUX/VBD reported/VERB/VBN
in/ADP/IN today/NOUN/NN 's/PART/POS New/PROPN/NNP
England/PROPN/NNP Journal/PROPN/NNP of/ADP/IN Medicine/PROPN/NNP

# POS Labelled Corpora

- examples:
  - Brown corpus (1961, $10^6$ words, different genre texts),
  - Wall Street Journal corpus (1989, $10^6$ words),
  - Switchboard corpus (1991, $2*10^6$ words, telephone conversations)

- slight differences in using POS tags (e.g. in corpora)
  - e.g.
    -- Brown, WSJ: to/TO for both uses of to (preposition: *go to the store;* infinitive: *too dangerous to swim*)
    -- Switchboard: Well/UH ,/, I/PRP ,/, I/PRP want/VBP to/TO go/VB
    to/IN a/DT restaurant/NN

# POS Labelled Corpora

- POS tag sets: pragmatic decisions:
  - Penn 45 is a subset of larger POS tagsets, leaving off syntactic information recoverable from a parse tree, e.g. in Penn, the tag IN is used for subordinating conjunctions
    **after/IN** spending/VBG a/DT day/NN at/IN the/DT beach/NN
    as well as prepositions:
    **after/IN** sunrise/NN
  - Penn 45 assumes tokenization of multipart words:
    a/DT New/NNP York/NNP City/NNP firm/NN (New York City as one word)

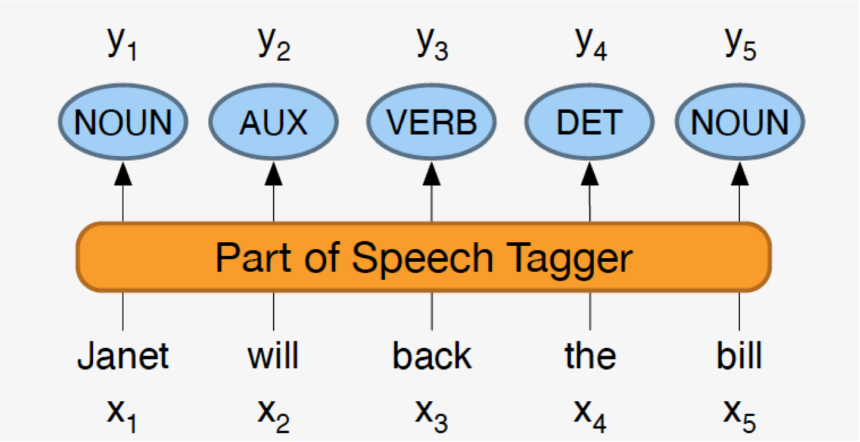
# POS Tagging

- After tokenization: POS tagging for each word: disambiguation task (*book a flight, read a book*)
  Not many words ambiguous but ambiguous words are among the most common tokens:

| Types: | | WSJ | | Brown | |
|---|---|---|---|---|---|
| Unambiguous | (1 tag) | 44,432 | (86%) | 45,799 | (85%) |
| Ambiguous | (2+ tags) | 7,025 | (14%) | 8,050 | (15%) |
| **Tokens:** | | | | | |
| Unambiguous | (1 tag) | 577,421 | (45%) | 384,349 | (33%) |
| Ambiguous | (2+ tags) | 711,780 | (55%) | 786,646 | (67%) |

- Most frequent POS tag (class) baseline: always predict the most frequent POS tag among the possible POS tags for an ambiguous word:
  on WSJ: accuracy: ≈0.92 ←→ state of the art: accuracy: ≈0.97

# POS Tagging




earnings growth took a **back/JJ** seat
a small building in the **back/NN**
a clear majority of senators **back/VBP** the bill
Dave began to **back/VB** toward the door
enable the country to buy **back/RP** debt
I was twenty-one **back/RB** then

# Named Entity Recognition

- Named Entity: Anything referred to by a proper name, often extended to temporal or numerical expressions

> 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**].

- Entity Types:

| Type | Tag | Sample Categories | Example sentences |
|---|---|---|---|
| People | PER | people, characters | **Turing** is a giant of computer science. |
| Organization | ORG | companies, sports teams | The **IPCC** warned about the cyclone. |
| Location | LOC | regions, mountains, seas | The **Mt. Sanitas** loop is in **Sunshine Canyon**. |
| Geo-Political Entity | GPE | countries, states, provinces | **Palo Alto** is raising the fees for parking. |
| Facility | FAC | bridges, buildings, airports | Consider the **Tappan Zee Bridge**. |
| Vehicles | VEH | planes, trains, automobiles | It was a classic **Ford Falcon**. |
| ... | | | |

# Named Entity Recognition

- Named entity recognition: finding spans of text that constitute NEs + classification

- categorial ambiguities:

| Name | Possible Categories |
|---|---|
| *Washington* | Person, Location, Political Entity, Organization, Vehicle |
| *Downing St.* | Location, Organization |
| *IRA* | Person, Organization, Monetary Instrument |
| *Louis Vuitton* | Person, Organization, Commercial Product |

[$_{\text{PER}}$ 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.
The [$_{\text{VEH}}$ Washington] had proved to be a leaky ship, every passage I made...

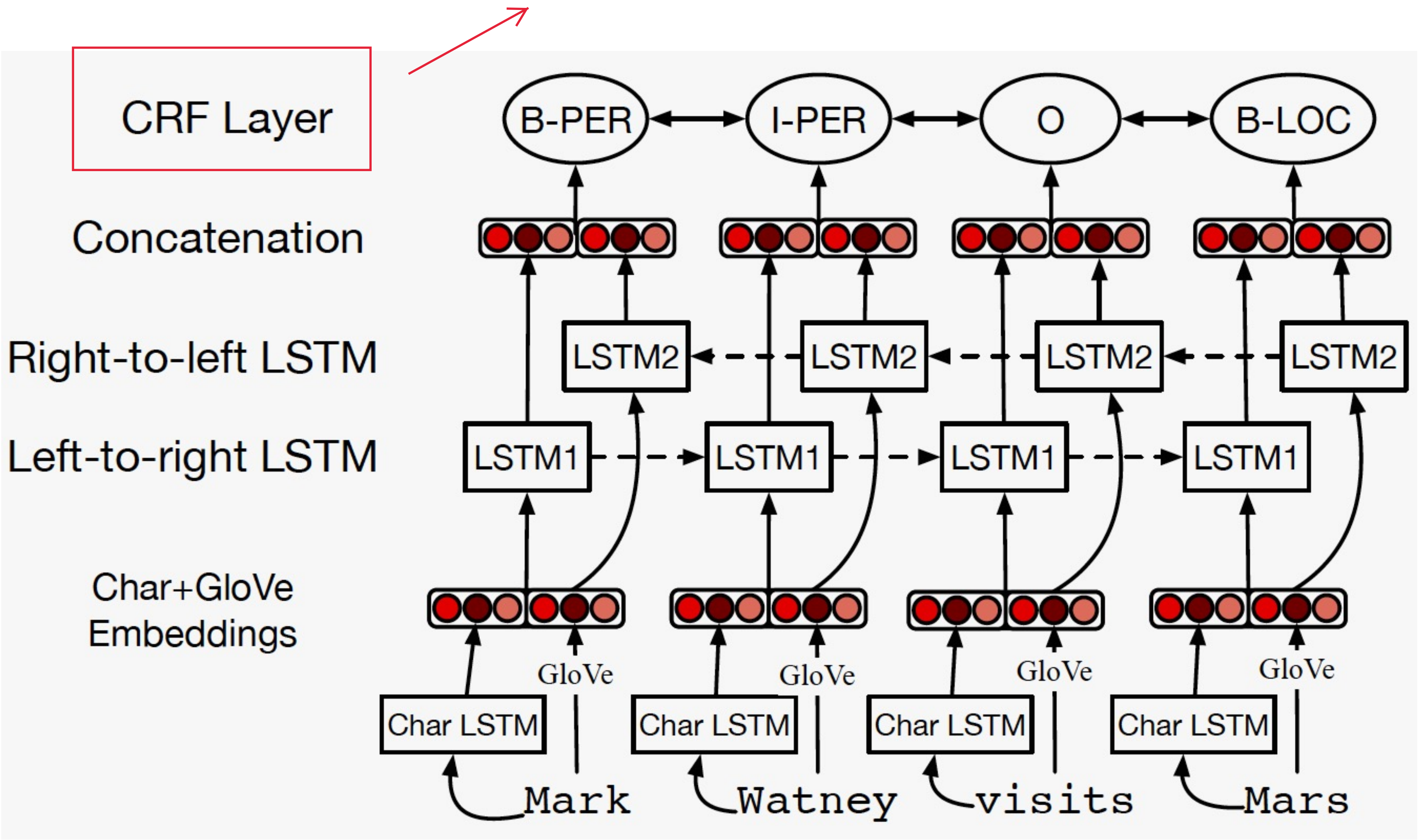
# NER as Sequence Labeling Task

- use supervised sequence classifier such as HMM or RNN, with BIO tagging (for n entity types → 2n+1 corresp. BIO classes) or IO tagging (n+1 corresp. IO classes) or BIOES tagging (E=end, S=one word span):

[$_{\text{PER}}$ **Jane Villanueva** ] of [$_{\text{ORG}}$ **United**] , a unit of [$_{\text{ORG}}$ **United Airlines Holding**] , said the fare applies to the [$_{\text{LOC}}$ **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 |

# NER as Sequence Labeling Task – NN Methods




(more on the details in the second half of the lecture)

# HMM for POS Tagging

If you do not know HMMs, study appendix A of Jurafsky and slide-set _A_HiddenMarkovModels

- States: tags; observations: words
- training on labelled data: MLE by counting for A and B separately (No Baum Welch necessary):

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

- POS-Tagging via Viterbi algorithm: find:

$$\hat{t}_{1:n} = \underset{t_1 \ldots t_n}{\operatorname{argmax}} P(t_1 \ldots t_n|w_1 \ldots w_n)$$

$$= \underset{t_1 \ldots t_n}{\operatorname{argmax}} P(w_1 \ldots w_n|t_1 \ldots t_n)P(t_1 \ldots t_n)$$

- First oder Markov assumptions for A and B:

$$P(w_1 \ldots w_n|t_1 \ldots t_n) \approx \prod_{i=1}^{n} P(w_i|t_i)$$

$$P(t_1 \ldots t_n) \approx \prod_{i=1}^{n} P(t_i|t_{i-1})$$

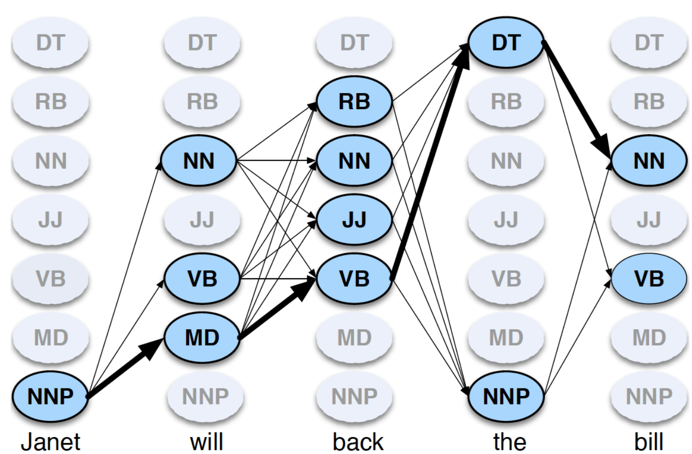
$$\hat{t}_{1:n} = \underset{t_1 \ldots t_n}{\operatorname{argmax}} P(t_1 \ldots t_n|w_1 \ldots w_n) \approx \underset{t_1 \ldots t_n}{\operatorname{argmax}} \prod_{i=1}^{n} \overbrace{P(w_i|t_i)}^{\text{emission}} \overbrace{P(t_i|t_{i-1})}^{\text{transition}}$$

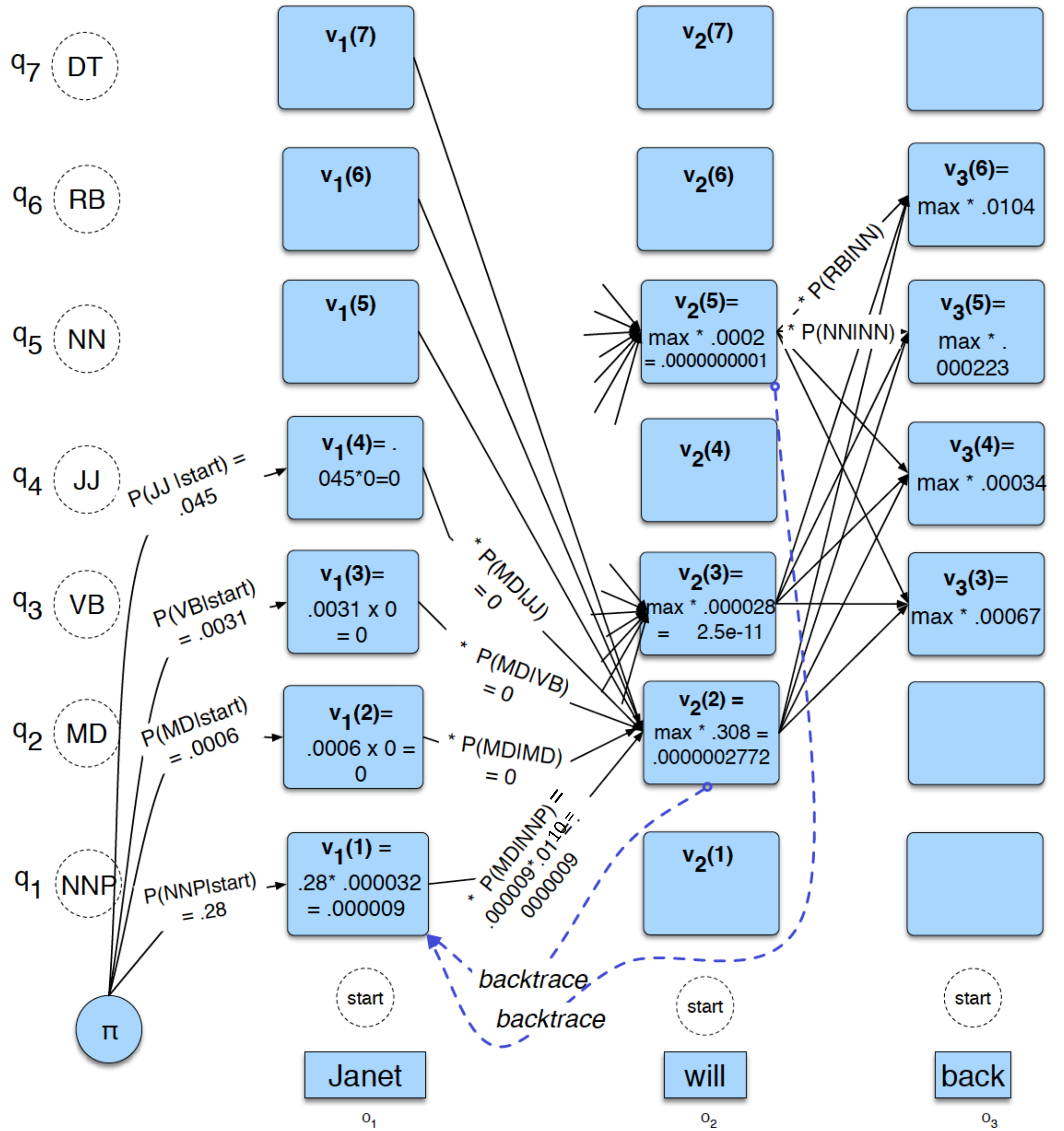
example: *Janet will back the bill* →
true POS tags:
Janet/NNP will/MD back/VB the/DT bill/NN

$$v_t(j) = \max_{q_{1:t-1}} P(q_{1:t-1}, q_t = j, o_{1:t} | \lambda)$$
$$= \max_{i=1}^{N} v_{t-1}(i)\, a_{ij}\, b_j(o_t)$$

| | NNP | MD | VB | JJ | NN | RB | DT |
|---|---|---|---|---|---|---|---|
| <s> | 0.2767 | 0.0006 | 0.0031 | 0.0453 | 0.0449 | 0.0510 | 0.2026 |
| NNP | 0.3777 | 0.0110 | 0.0009 | 0.0084 | 0.0584 | 0.0090 | 0.0025 |
| MD | 0.0008 | 0.0002 | 0.7968 | 0.0005 | 0.0008 | 0.1698 | 0.0041 |
| VB | 0.0322 | 0.0005 | 0.0050 | 0.0837 | 0.0615 | 0.0514 | 0.2231 |
| JJ | 0.0366 | 0.0004 | 0.0001 | 0.0733 | 0.4509 | 0.0036 | 0.0036 |
| NN | 0.0096 | 0.0176 | 0.0014 | 0.0086 | 0.1216 | 0.0177 | 0.0068 |
| RB | 0.0068 | 0.0102 | 0.1011 | 0.1012 | 0.0120 | 0.0728 | 0.0479 |
| DT | 0.1147 | 0.0021 | 0.0002 | 0.2157 | 0.4744 | 0.0102 | 0.0017 |

| | Janet | will | back | the | bill |
|---|---|---|---|---|---|
| NNP | 0.000032 | 0 | 0 | 0.000048 | 0 |
| MD | 0 | 0.308431 | 0 | 0 | 0 |
| VB | 0 | 0.000028 | 0.000672 | 0 | 0.000028 |
| JJ | 0 | 0 | 0.000340 | 0.000097 | 0 |
| NN | 0 | 0.000200 | 0.000223 | 0.000006 | 0.002337 |
| RB | 0 | 0 | 0.010446 | 0 | 0 |
| DT | 0 | 0 | 0 | 0.506099 | 0 |



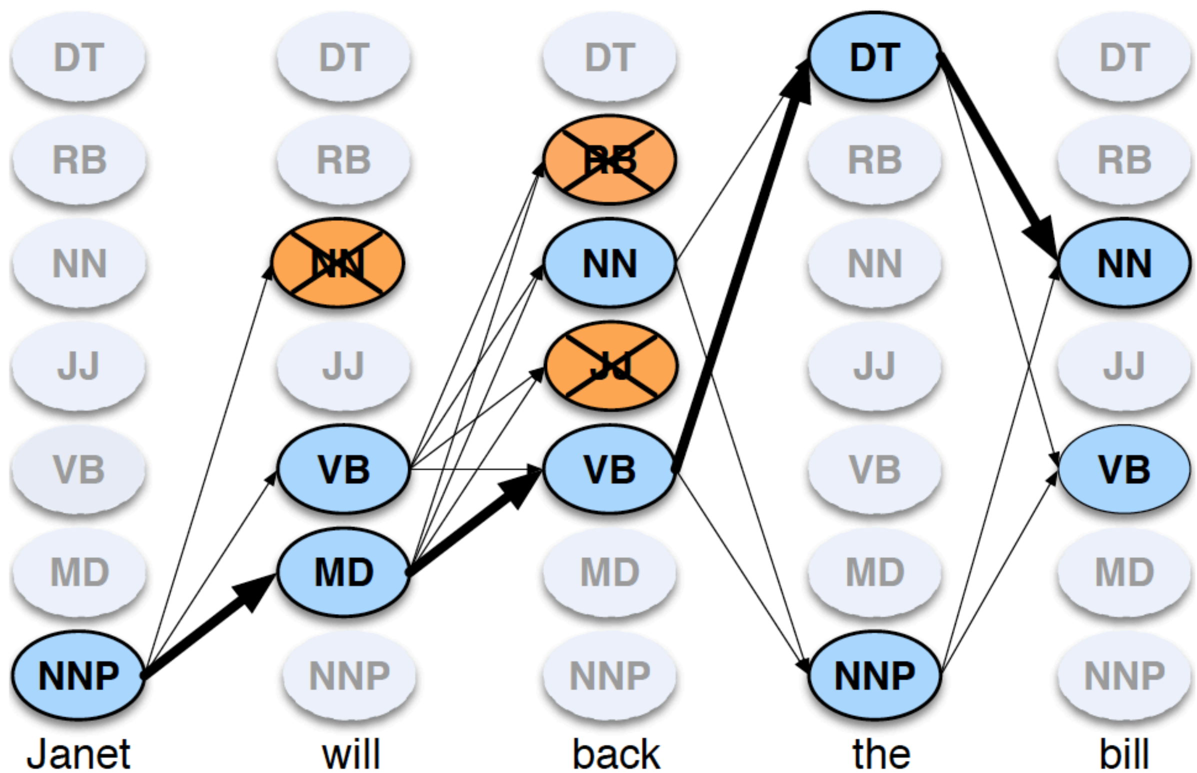



# Beam Search

- no of states N large → Viterbi ($O(N^2T)$) inefficient →
- instead of keeping all N=45 possibilities at each column, just concentrate on the $\beta$ most probable ones (prune the possible hidden sequence tree);

becasuse impossible to calcuate

- $\beta$ : beam width



# Conditional Random Fields (CRFs)

- Problems of HMMs: unknown words. Possible remedies:
  -- use morphological features: ___*-ed* → VBN or VBD more likely
  -- deviate from Markov first order (use previous word or next word)

- Generative Models (like HMM): incorporate those features: cumbersome → switch to discriminative model (compare Naïve Bayes → Logistic Regression): new features easy to include in conditioning side

- discriminative sequence processing model: linear chain CRF

tag-sequence

word-sequence

HMM:

$$\begin{aligned}\hat{Y} &= \operatorname*{argmax}_{Y} p(Y|X) \\ &= \operatorname*{argmax}_{Y} p(X|Y)p(Y) \\ &= \operatorname*{argmax}_{Y} \prod_i p(x_i|y_i) \prod_i p(y_i|y_{i-1})\end{aligned}$$

CRF:

$$\hat{Y} = \operatorname*{argmax}_{Y \in \mathcal{Y}} P(Y|X)$$

# CRFs

$$p(Y|X) = \frac{\exp\left(\sum_{k=1}^{K} w_k F_k(X,Y)\right)}{\sum_{Y' \in \mathcal{Y}} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y')\right)}$$

$$p(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y)\right)$$

$$Z(X) = \sum_{Y' \in \mathcal{Y}} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y')\right)$$

$$\underbrace{F_k(X,Y)}_{\text{global feature}} = \sum_{i=1}^{n} \underbrace{f_k(y_{i-1}, y_i, X, i)}_{\text{local feature}}$$

# CRFs

$$p(Y|X) = \frac{\exp\left(\sum_{k=1}^{K} w_k F_k(X,Y)\right)}{\sum_{Y' \in \mathcal{Y}} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y')\right)}$$

$$p(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y)\right)$$

$$Z(X) = \sum_{Y' \in \mathcal{Y}} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y')\right)$$

$$F_k(X,Y) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i)$$

restriction to current and previous tags:
„linear chain“ → variants of HMM
algorithms (e.g. Viterbi) can be used

# CRF Features for POS Tagging

- any feature constructed from $(y_{i-1}, y_i, X, i)$ e.g.

$$\mathbb{1}\{x_i = \textit{the},\ y_i = \text{DET}\}$$
$$\mathbb{1}\{y_i = \text{PROPN},\ x_{i+1} = \textit{Street},\ y_{i-1} = \text{NUM}\}$$
$$\mathbb{1}\{y_i = \text{VERB},\ y_{i-1} = \text{AUX}\}$$

- e.g. using feature templates such as

$$\langle y_i, x_i \rangle, \langle y_i, y_{i-1} \rangle, \langle y_i, x_{i-1}, x_{i+2} \rangle$$

example : for Janet/NNP will/MD back/VB the/DT bill/NN and $x_i$=*back:*

$f_{3743}$: $y_i$ = VB and $x_i$ = back
$f_{156}$: $y_i$ = VB and $y_{i-1}$ = MD
$f_{99732}$: $y_i$ = VB and $x_{i-1}$ = will and $x_{i+2}$ = bill

# CRF Features for POS Tagging

- useful for unknown words: word shape features: x: letter; X: uppercase letter; d: number; punctuation.
  examples: I.M.F. → X.X.X   DC10-30 → XXdd-dd

- useful for unknown words: prefix- or suffix-features:

  $x_i$ contains a particular prefix (perhaps from all prefixes of length $\leq 2$)
  $x_i$ contains a particular suffix (perhaps from all suffixes of length $\leq 2$)
  $x_i$'s word shape
  $x_i$'s short word shape

example: *well-dressed*:

$$\text{prefix}(x_i) = \texttt{w}$$
$$\text{prefix}(x_i) = \texttt{we}$$
$$\text{suffix}(x_i) = \texttt{ed}$$
$$\text{suffix}(x_i) = \texttt{d}$$
$$\text{word-shape}(x_i) = \texttt{xxxx-xxxxxxx}$$
$$\text{short-word-shape}(x_i) = \texttt{x-x}$$

# CRF Features for NER

identity of $w_i$, identity of neighboring words
embeddings for $w_i$, embeddings for neighboring words
part of speech of $w_i$, part of speech of neighboring words
presence of $w_i$ in a **gazetteer** list of names in spesific domain
$w_i$ contains a particular prefix (from all prefixes of length $\leq 4$)
$w_i$ contains a particular suffix (from all suffixes of length $\leq 4$)
word shape of $w_i$, word shape of neighboring words
short word shape of $w_i$, short word shape of neighboring words
gazetteer features

**Figure 8.15** Typical features for a feature-based NER system.

| Words | POS | Short shape | Gazetteer | BIO Label |
|---|---|---|---|---|
| Jane | NNP | Xx | 0 | B-PER |
| Villanueva | NNP | Xx | 1 | I-PER |
| of | IN | x | 0 | O |
| United | NNP | Xx | 0 | B-ORG |
| Airlines | NNP | Xx | 0 | I-ORG |
| Holding | NNP | Xx | 0 | I-ORG |
| discussed | VBD | x | 0 | O |
| the | DT | x | 0 | O |
| Chicago | NNP | Xx | 1 | B-LOC |
| route | NN | x | 0 | O |
| . | . | . | 0 | O |

**Figure 8.16** Some NER features for a sample sentence, assuming that Chicago and Villanueva are listed as locations in a gazetteer. We assume features only take on the values 0 or 1, so the first POS feature, for example, would be represented as $\mathbb{1}\{\text{POS} = \text{NNP}\}$.

# CRF Inference & Training

$$
\begin{aligned}
\hat{Y} &= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}}\, P(Y|X) \\
&= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}}\, \frac{1}{Z(X)} \exp\left( \sum_{k=1}^{K} w_k F_k(X, Y) \right) \\
&= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}}\, \exp\left( \sum_{k=1}^{K} w_k \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i) \right) \\
&= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{k=1}^{K} w_k \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i) \\
&= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{k=1}^{K} w_k f_k(y_{i-1}, y_i, X, i)
\end{aligned}
$$

# CRF Inference & Training

- For HMM we had:

$$\hat{t}_{1:n} = \underset{t_1 \ldots t_n}{\operatorname{argmax}} P(t_1 \ldots t_n | w_1 \ldots w_n) \approx \underset{t_1 \ldots t_n}{\operatorname{argmax}} \prod_{i=1}^{n} \overbrace{P(w_i|t_i)}^{\text{emission}} \overbrace{P(t_i|t_{i-1})}^{\text{transition}}$$

Viterbi:

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i)\, P(s_j|s_i)\, P(o_t|s_j) \quad 1 \le j \le N, 1 < t \le T$$

$$= \max_{i=1}^{N} v_{t-1}(i)\, a_{ij}\, b_j(o_t); \quad 1 \le j \le N, 1 < t \le T$$

- HMM Viterbi formulated in terms of tags t (also denoted as states s) as y and outputs w (also denoted as o) as x:

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i) P\left(y_j \middle| y_i\right) P(x_t|y_j) \quad 1 \le j \le N, 1 < t \le T$$

- Linear chain CRF version of Viterbi

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i) \sum_{k=1}^{K} w_k f_k(y_{t-1}, y_t, X, t) \quad 1 \le j \le N, 1 < t \le T$$

# Practical Variation for POS and NER

- Combine rule-based with ML-based:

  1. First, use high-precision rules to tag unambiguous entity mentions.
  2. Then, search for substring matches of the previously detected names.
  3. Use application-specific name lists to find likely domain-specific mentions.
  4. Finally, apply supervised sequence labeling techniques that use tags from previous stages as additional features.

- POS-Tagging for morphologically rich laguages: use morphologically rich POS tags to be able to deal with unseen variants of words (→ 4-10 times larger tag-set)

  1. Yerdeki **izin** temizlenmesi gerek. — iz + Noun+A3sg+Pnon+Gen
     **The trace** on the floor should be cleaned.
  2. Üzerinde parmak **izin** kalmiş — iz + Noun+A3sg+P2sg+Nom
     **Your** finger **print** is left on (it).
  3. Içeri girmek için **izin** alman gerekiyor. — izin + Noun+A3sg+Pnon+Nom
     You need **permission** to enter.

# Bibliography

(1) Dan Jurafsky and James Martin: Speech and Language Processing (3rd ed. draft, version Jan2022); Online: https://web.stanford.edu/~jurafsky/slp3/ (URL, Oct 2022) (this slide-set is especially based on chapter 8)

# Recommendations for Studying

- minimal approach:

  work with the slides and understand their contents! Think beyond instead of merely memorizing the contents

- standard approach:

  minimal approach + read the corresponding pages in Jurafsky [1]

- interested students

  == standard approach