

Processamento de Linguagem Natural

Extração de Informação

Prof. Luciano Barbosa &
Prof. Johny Moreira
{luciano, jms5}@cin.ufpe.br

Objetivo

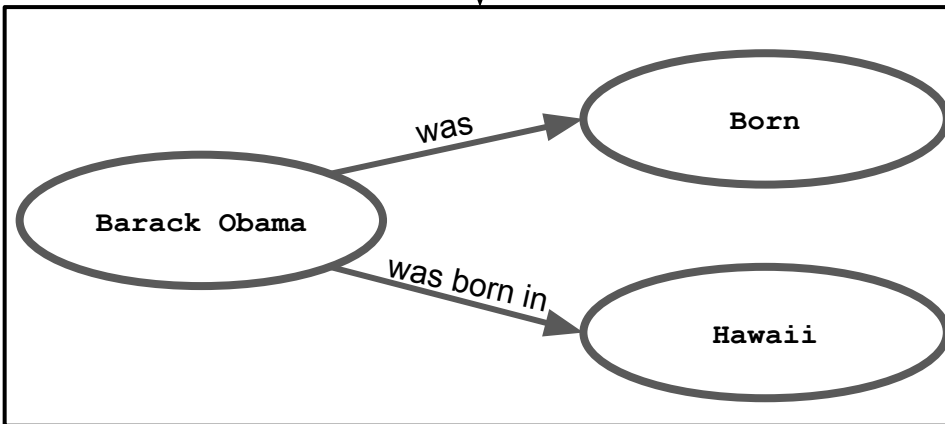
Extrair estrutura a partir de dados não estruturados

Unstructured Text

Barack Obama was born in Hawaii.

Information
Extraction

Structured Text



Rotulagem Sequencial

- ❖ Objetivo: atribuir um dado rótulo a cada palavra de um sentença
- ❖ Rótulos dependem de outras palavras da sequência (não é i.i.d)

Algumas Tarefas...

- ❖ Named Entity Recognition
- ❖ Rotulagem de Papel Semântico
- ❖ Part-of-Speech Tagging (Rotulagem de Classe Gramatical)
- ❖ Bioinformática

Named Entity Recognition

- ❖ Identificar nomes de pessoas, locais etc no texto

people

organizations

places

Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

- ❖ Extrair partes de informação relevante para uma dada aplicação

make

model

year

mileage

price

For sale, 2002 Toyota Prius, 20,000 mi, \$15K or best offer.
Available starting July 30, 2006.

Rotulagem de Papel Semântico

Determina o papel semântico de cada noun phrase que é argumento do verbo

agent patient source destination instrument

John drove Mary from Austin to Dallas in his Toyota Prius.

The hammer broke the window.

Bioinformática

Rotular sequências genéticas

exon

intron

AGCTAACGTT**CGATACG****GATTACAGCCT**

Como identificar cada um desses rótulos?

Part-of-Speech Tagging

Named Entity Recognition

Part-of-speech Tagging

- ❖ Atribuir a classe gramatical a cada palavra de uma sentença (substantivo, adjetivo, verbo etc)
- ❖ Útil para tarefa de desambiguação: palavras podem ter mais de uma classe gramatical
Ex: book, that etc
- ❖ Classe mais frequente da palavra já tem alta acurácia

John	saw	the	saw	and	decided	to	take	it	to	the	table.
PN	V	Det	N	Con	V	Part	V	Pro	Prep	Det	N

Part-of-speech Tagging

- ❖ Pequena proporção das palavras possui mais de uma classe
- ❖ Palavras com mais de uma classe são mais frequentes

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%)

Figure 8.4 Tag ambiguity in the Brown and WSJ corpora (Treebank-3 45-tag tagset).

Part-of-speech 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

Part-of-speech Tagging

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

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).

Tipos de POS

❖ Closed class:

- Preposições e pronomes
- Tendem a ser curtos
- Alta frequência

❖ Open class:

- Substantivos, verbos, adjetivos e advérbios
- Constantemente sendo criados

POS no Penn Treebank

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

Figure 8.2 Penn Treebank part-of-speech tags.

Named Entity Recognition (Information Extraction)

- ❖ Named entity: tudo que se refere a um nome próprio (regra geral)

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	Mt. Sanitas is in Sunshine Canyon .
Geo-Political Entity	GPE	countries, states	Palo Alto is raising the fees for parking.

Figure 8.5 A list of generic named entity types with the kinds of entities they refer to.

- ❖ Pode ser qualquer entidade: produto, doenças etc
- ❖ Usado em Natural Language Understanding: Q&A, chatbot
- ❖ Dificuldades:
 - Encontrar o pedaço do texto que contém a entidade
 - Ambiguidade: JFK (pessoa ou aeroporto)

BIO Tagging

Convenção para rotulagem de sequência

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

Figure 8.7 NER as a sequence model, showing IO, BIO, and BIOES taggings.

Modelos de Rotulagem Sequencial

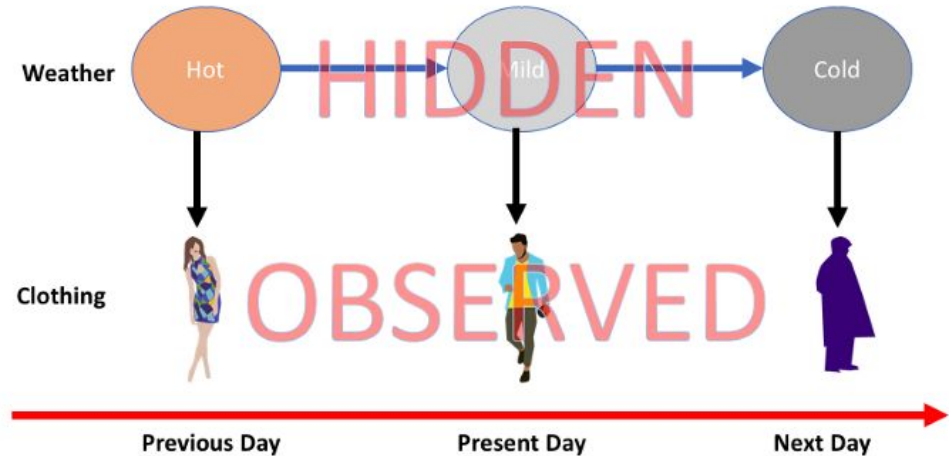
Hidden Markov Model (HMM)

Conditional Random Fields (CRF)

Recurrent Neural Networks (RNN)

Hidden Markov Model para POS Tagging

- ❖ Modelo probabilístico sequencial
- ❖ Computa a probabilidade para possíveis sequências de rótulos
- ❖ Escolhe a melhor sequência
- ❖ Rótulos estão escondidos (hidden)
- ❖ Observa palavras
- ❖ Inferir rótulos (ex. POS) da sequência de palavras

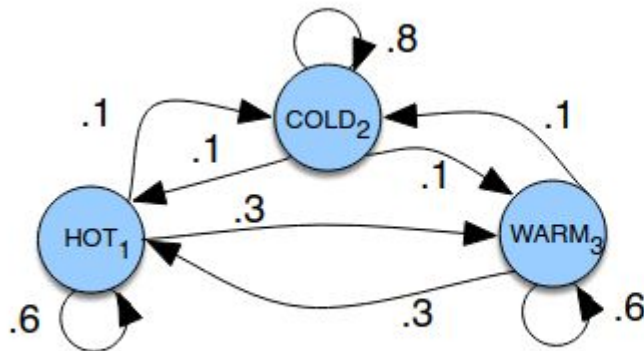


Markov Chain

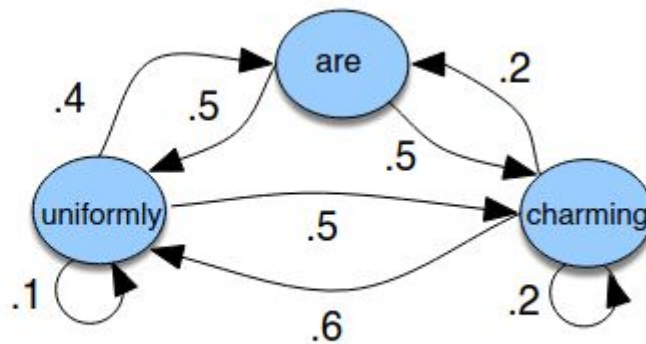
- ❖ Probabilidade do próximo estado só vai depender do estado anterior
- ❖ Quando prevendo o rótulo futuro, o passado não importa. Somente o estado presente.

Markov Assumption

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

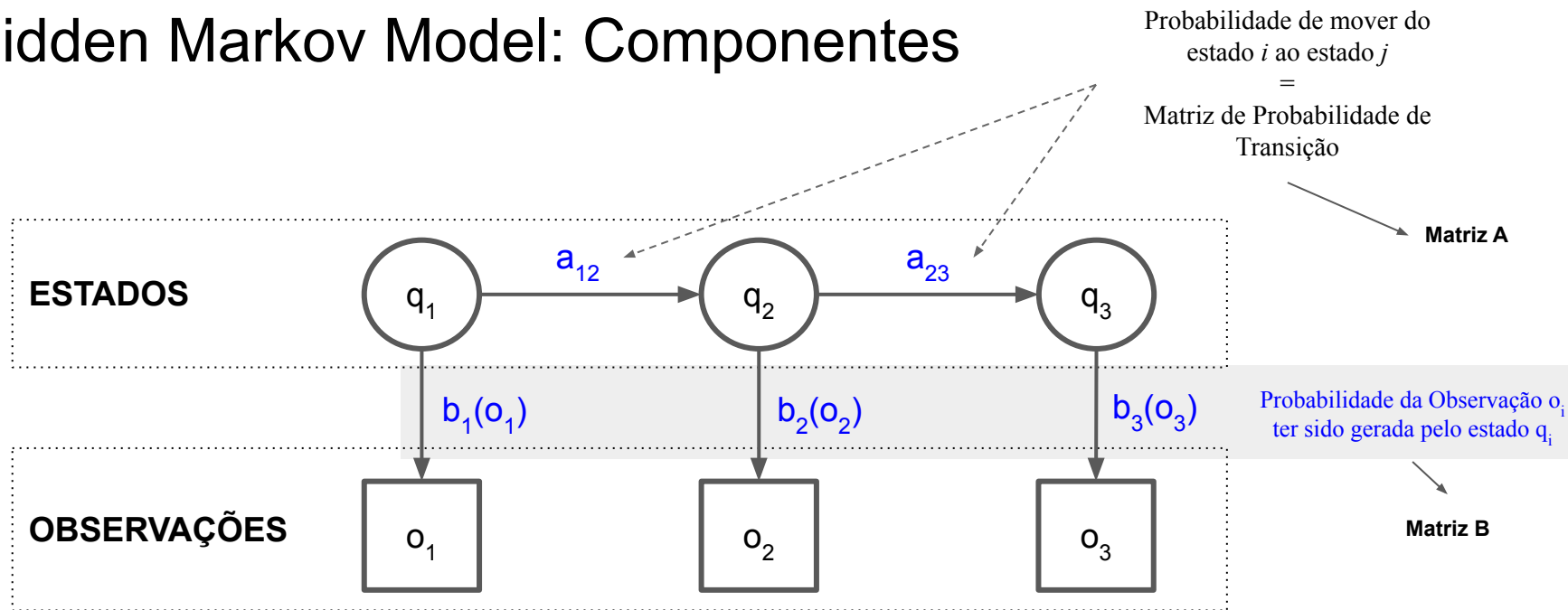


(a)



(b)

Hidden Markov Model: Componentes



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

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

HMM Tagger: Componentes

Matriz A: probabilidades de transição das tags

A probabilidade de uma tag ocorrer observando-se uma tag anterior

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

Contagem

Exemplo No corpus WSJ, **MD** ocorre 13.124 no corpus sendo que 10.471 vezes a tag **MD** aparece seguida pela tag **VB**. Logo, a probabilidade de termos uma tag **MD** seguida de **VB** é:

$$P(VB|MD) = \frac{C(MD, VB)}{C(MD)} = \frac{10471}{13124} = .80$$

MD: Modal
VB: Verb

HMM Tagger: Componentes

Matriz B: probabilidades de uma palavra associada a uma tag

A probabilidade de uma
palavra dado que foi
observada uma tag

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

Contagem

Exemplo Das 13.124 ocorrências da tag MD no corpus WSJ, a tag está associada 4046 vezes à palavra “will”. Logo, a probabilidade de termos a palavra “will” associada à tag MD é dada por:

$$P(will|MD) = \frac{C(MD, will)}{C(MD)} = \frac{4046}{13124} = .31$$

HMM Tagger: Exemplo

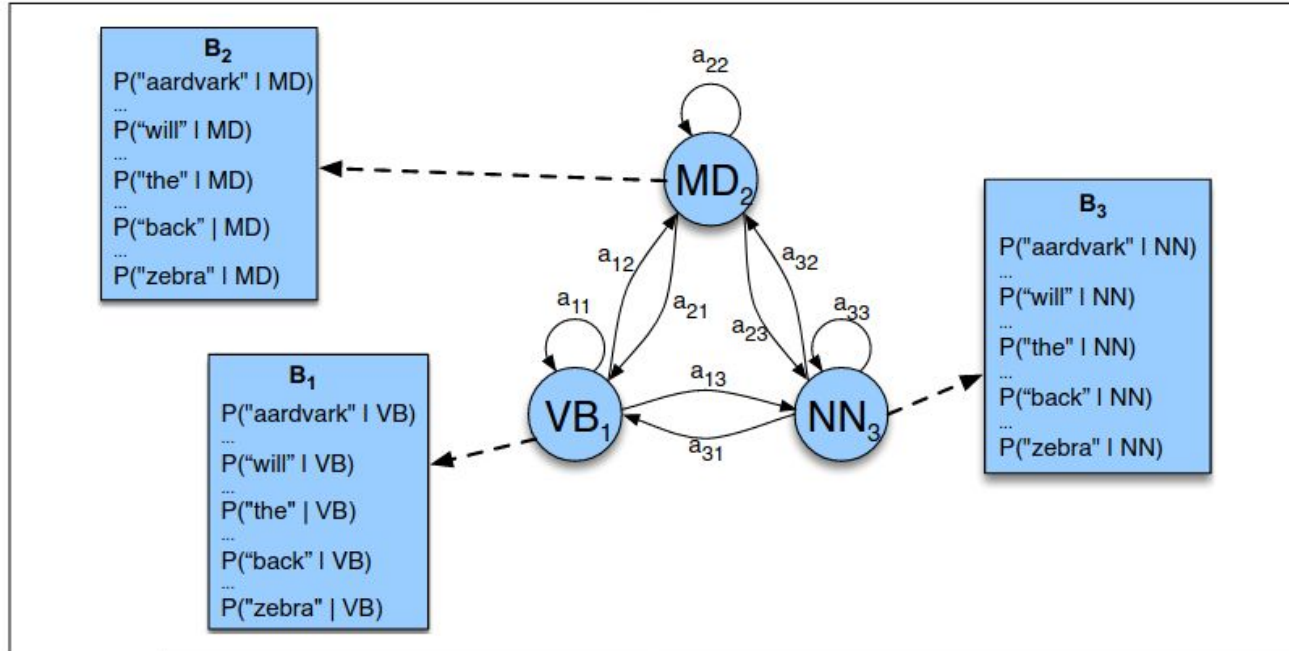



Figure 8.9 An illustration of the two parts of an HMM representation: the A transition probabilities used to compute the prior probability, and the B observation likelihoods that are associated with each state, one likelihood for each possible observation word.

HMM Decoding

- ❖ Dadas as **matrizes A e B** como entrada, assim como a **sequência de palavras** (observações), o objetivo é encontrar a sequência de tags mais prováveis.
- ❖ Teorema de Bayes
- ❖ Suposições:
 1. A probabilidade de uma palavra aparecer na sequência é independente da vizinhança e depende somente da sua tag;
 2. A probabilidade de uma tag depende somente da tag anterior

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


Matriz B **Matriz A**

HMM: O Algoritmo Viterbi

O valor de cada célula é computado recursivamente obtendo o caminho mais provável

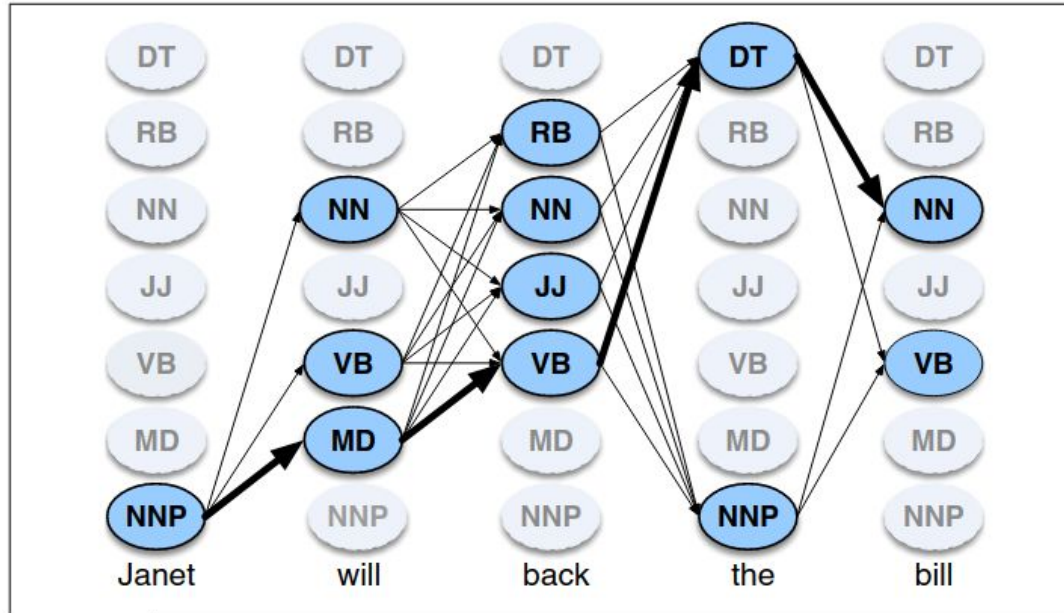


Figure 8.11 A sketch of the lattice for *Janet will back the bill*, showing the possible tags (q_i) for each word and highlighting the path corresponding to the correct tag sequence through the hidden states. States (parts of speech) which have a zero probability of generating a particular word according to the B matrix (such as the probability that a determiner DT will be realized as *Janet*) are greyed out.

HMM: O Algoritmo Viterbi

Matriz A

	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

Figure 8.12 The A transition probabilities $P(t_i|t_{i-1})$ computed from the WSJ corpus without smoothing. Rows are labeled with the conditioning event; thus $P(VB|MD)$ is 0.7968.

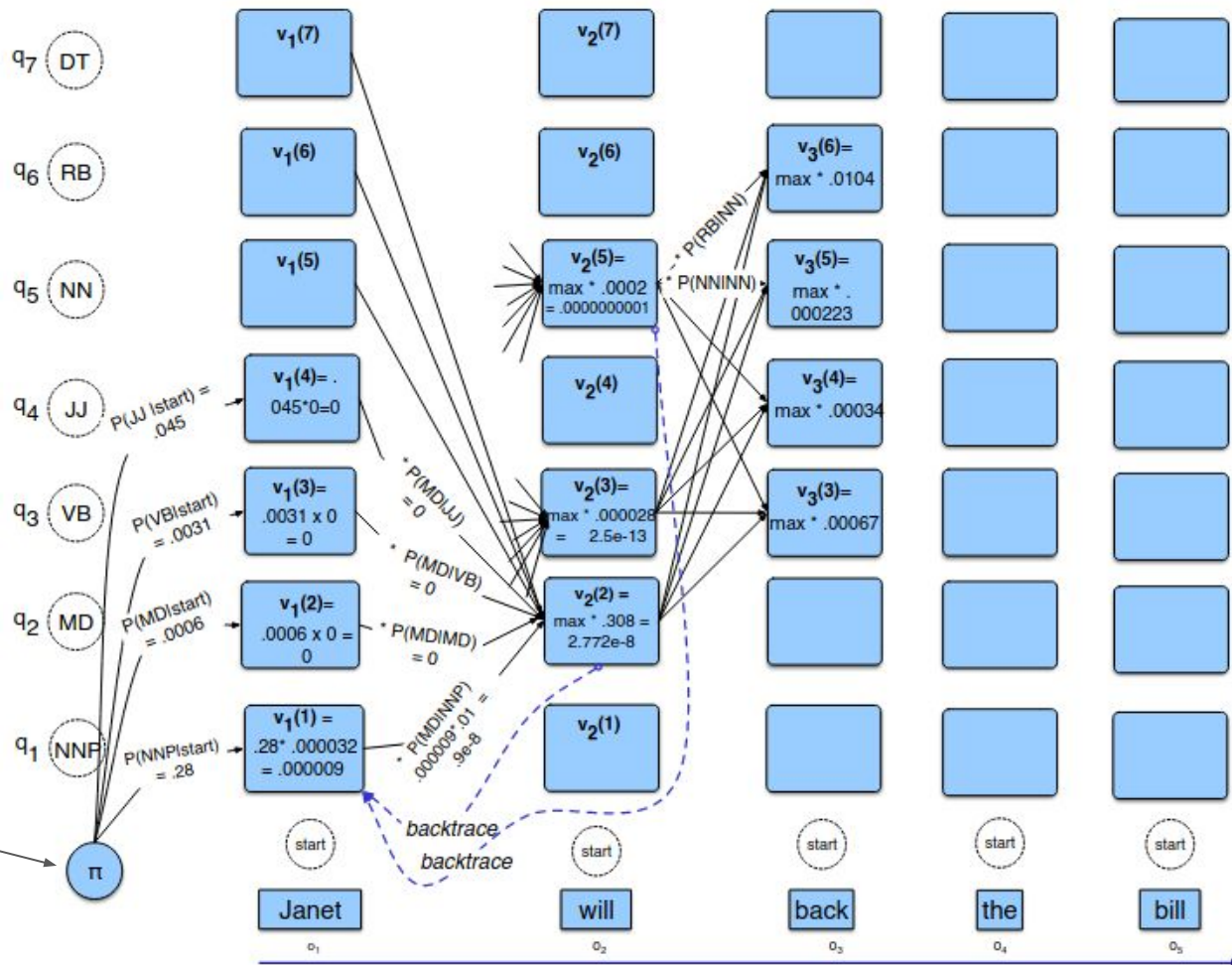
Matriz B

	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	0
NN	0	0.000200	0.000223	0	0.002337
RB	0	0	0.010446	0	0
DT	0	0	0	0.506099	0

Figure 8.13 Observation likelihoods B computed from the WSJ corpus without smoothing, simplified slightly.

HMM: O Algoritmo Viterbi

Distribuição de probabilidade inicial. Na Matriz A é dado por $\langle s \rangle$



HMM

❖ É um modelo generativo:

- modela como os dados foram gerados e depois aplica o que foi aprendido para classificar cada item da sequência
- Modela a distribuição de probabilidade conjunta

❖ Modelo útil e poderoso

❖ Problemas:

- Precisa de muitos dados para alcançar boa acurácia
- Dificuldade nas tarefas NLP: a existência de palavras desconhecidas
 - Nomes próprios ou acrônimos por exemplo
- Limitação das features:
 - Não diferencia maiúsculas ou minúsculas
 - Não considera o contexto anterior da palavra

Conditional Random Fields (CRF)

- ❖ Modelo sequencial discriminativo baseado em modelo log-linear
- ❖ Aplica regressão logística a sequências
- ❖ É um modelo discriminativo:
 - modela as fronteiras de decisão entre as classes
 - Modela a distribuição de probabilidade condicional
- ❖ Não supõe que os itens são independentes
- ❖ A cada passo computa funções log-lineares sobre um conjunto de features
- ❖ Dada uma entrada X , computa a probabilidade de uma sequência inteira Y

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

CRF: Features

Número de características

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

Peso para cada característica

Feature Global:
Mapeia as sequências de entrada e saída inteiras para um vetor de características

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

Local Features

X: toda a sequência de entrada (ou uma parte dela)
 y_i : token atual da saída
 y_{i-1} : token anterior
i: posição atual

CRF: Features (exemplo)

Word: “*well-dressed*”

$\text{prefix}(x_i) = w$

$\text{prefix}(x_i) = we$

$\text{suffix}(x_i) = ed$

$\text{suffix}(x_i) = d$

$\text{word-shape}(x_i) = xxxx-xxxxxxx$

$\text{short-word-shape}(x_i) = x-x$

Descrição das Features

x_i : contains a particular prefix (perhaps from all prefixes of length ≤ 2)

x_i : contains a particular suffix (perhaps from all suffixes of length ≤ 2)

x_i 's word shape

x_i 's short word shape

CRF: Features

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 $\mathbb{1}\{\text{POS} = \text{NNP}\}$.

Recurrent Neural Networks (RNN)

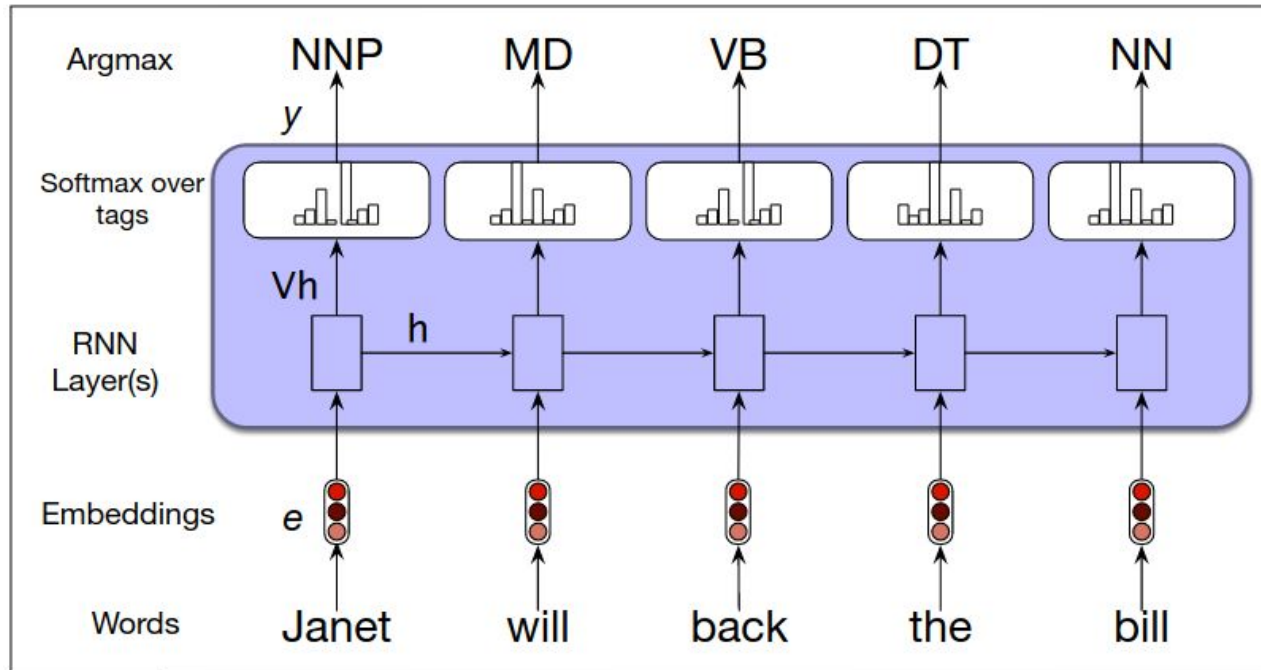


Figure 9.7 Part-of-speech tagging as sequence labeling with a simple RNN. Pre-trained word embeddings serve as inputs and a softmax layer provides a probability distribution over the part-of-speech tags as output at each time step.

Extração de Relações

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].

Relations

United is a unit of UAL

$PartOf = \{\langle a, b \rangle, \langle c, d \rangle\}$

American is a unit of AMR

Tim Wagner works for American Airlines

$OrgAff = \{\langle c, e \rangle\}$

United serves Chicago, Dallas, Denver, and San Francisco

$Serves = \{\langle a, f \rangle, \langle a, g \rangle, \langle a, h \rangle, \langle a, i \rangle\}$

Extração de Relações

Relations	Types	Examples
Physical-Located	PER-GPE	He was in Tennessee
Part-Whole-Subsidiary	ORG-ORG	XYZ , the parent company of ABC
Person-Social-Family	PER-PER	Yoko 's husband John
Org-AFF-Founder	PER-ORG	Steve Jobs , co-founder of Apple...

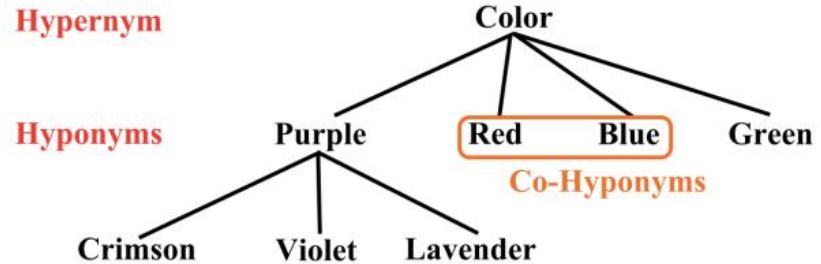
Figure 17.2 Semantic relations with examples and the named entity types they involve.

Extração de Relações Baseada em Padrões

- ❖ Padrões léxicos-sintáticos
- ❖ Hearst Patterns para extração de hipônimos

são palavras de sentido específico, ou seja, palavras cujos significados são hierarquicamente mais específicos do que de outras

*“Agar is a substance prepared from a mixture of red algae, such as **Gelidium**, for laboratory or industrial use.”*



NP_0 such as $NP_1\{, NP_2 \dots, (and|or)NP_i\}, i \geq 1$

$\forall NP_i, i \geq 1, \text{hyponym}(NP_i, NP_0)$

$\text{hyponym}(\text{Gelidium}, \text{red algae})$

Extração de Relações Baseada em Padrões

Hiperônimos são palavras de sentido genérico, ou seja, palavras cujos significados são mais abrangentes do que os hipônimos:

*Animais é hiperônimo de cachorro e cavalo.
Legume é hiperônimo de batata e cenoura.
Galáxia é hiperônimo de estrelas e planetas.*

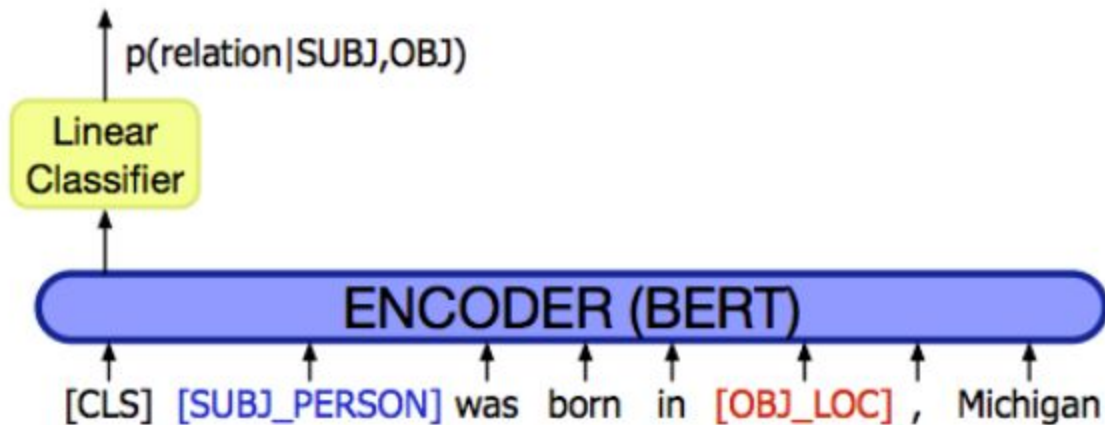
NP {, NP}* {,} (and or) other NP _H	temples, treasures, and other important civic buildings
NP _H such as {NP,*} {(or and)} NP	red algae such as Gelidium
such NP _H as {NP,*} {(or and)} NP	such authors as Herrick, Goldsmith, and Shakespeare
NP _H {,} including {NP,*} {(or and)} NP	common-law countries , including Canada and England
NP _H {,} especially {NP,*} {(or and)} NP	European countries , especially France, England, and Spain

Figure 17.5 Hand-built lexico-syntactic patterns for finding hypernyms, using { } to mark optionality (Hearst 1992a, Hearst 1998).

Extração de Relação Baseada em ML

- ❖ Definem-se as relações e entidades a serem extraídas
- ❖ Anotam-se exemplos para treinamento
- ❖ Features:
 - BOW e bigramas nas entidades
 - American, Airlines, Tim, Wagner, American Airlines, Tim Wagner
 - Palavras ou bigramas ao redor
 - Tipos das entidades
 - Número de entidades entre as entidades candidatas
 - Part-of-speech

Extração de Relação Baseada em Redes Neurais



Aula Prática

Extração de Informação com RNN

Extração de Informação com CRF

Referências

Dan Jurafsky, James H. Martin. Speech and Language Processing. (3rd ed. Draft). 2021.
Disponível em: <<https://web.stanford.edu/~jurafsky/slp3/8.pdf>>. Capítulo 8. Acesso em: 01 Setembro de 2022.

Dan Jurafsky, James H. Martin. Speech and Language Processing. (3rd ed. Draft). 2021.
Disponível em: <<https://web.stanford.edu/~jurafsky/slp3/17.pdf>>. Capítulo 17. Acesso em: 01 Setembro de 2022.