Information Extraction

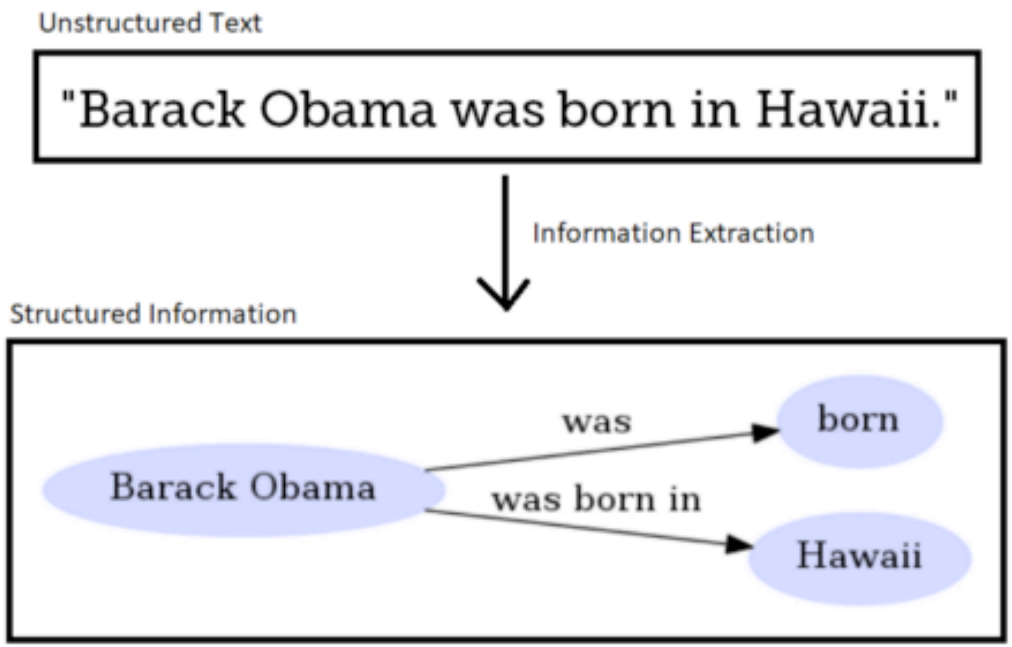
Luciano Barbosa

Centro de Informática
UFPE

Universidade Federal de Pernambuco

Objetivo

- Extrair estrutura a partir de dados não estruturados



Sequence Labeling

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

Part Of Speech Tagging

- Objetivo: atribuir a classe gramatical a cada palavra de uma sentença (substantivo, adjetivo, verbo etc)

- Útil para parsing sintático e desambiguação de palavras

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

Sequence Tagging

- Objetivo: atribuir rótulos a palavras de uma sentença (pessoa, local, empresa)
- Named entity recognition: identifica nomes de pessoas, locais 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.

Semantic Role Labeling

- 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

extron intron

– AGCTAACGTTCGATACGGATTACAGCCT

Principais POS (Inglês)

	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

Tipos de POS

- Closed class:
  - Preposições e pronomes
  - Tendem a ser curtos
  - Alta frequência
- Open class:
  - Substativos, verbos, adjetivos e advérbios
  - Constantemente sendo criados

POS no Penn Treebank

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*

Exemplos

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

Part-of-Speech Tagging

- Atribuir a classe gramatical a cada palavra de uma sentença (substantivo, adjetivo, verbo etc)
- Tarefa de desambiguação: palavras podem ter mais de uma class gramatical
  - Ex: book, that etc
- Classe mais frequente: 92% de acurácia
- Estado da arte: 97%

		WSJ		Brown	
Types:					
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%**)

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.

- Pode ser qualquer entidade: produto, doenças etc
- 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

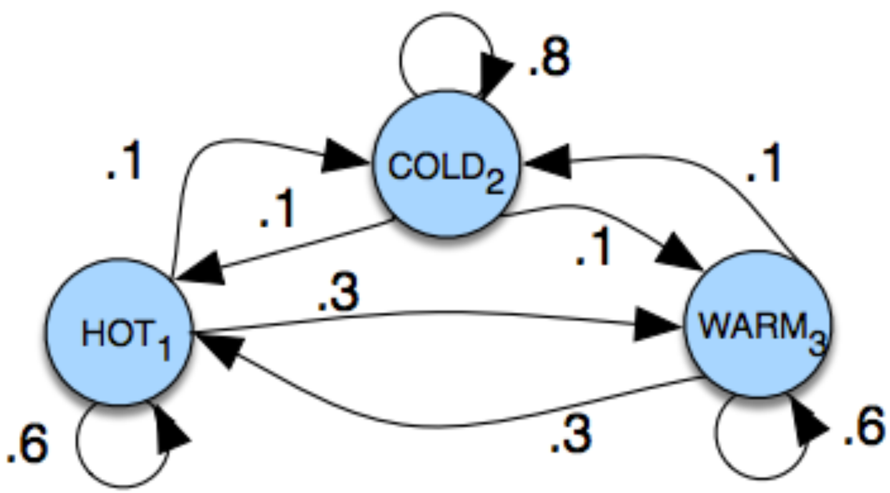
HMM POS Tagging

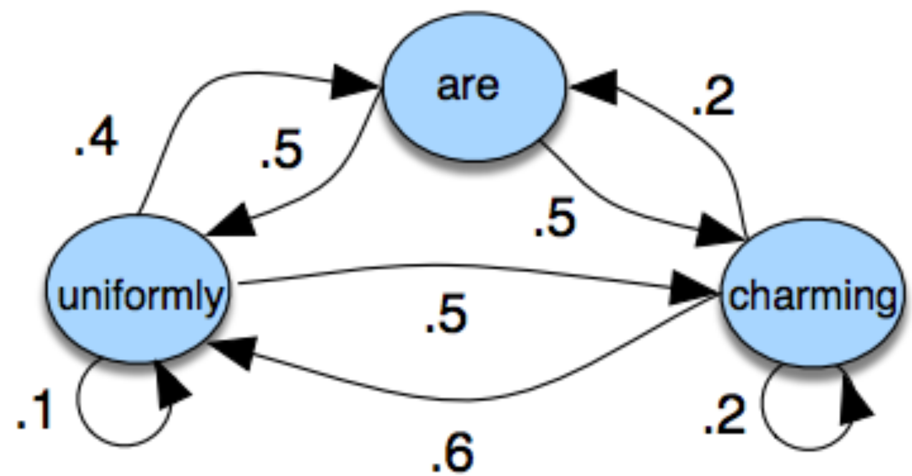
- Modelo probabilístico sequencial
- Computa a probabilidade para possíveis sequências de rótulos
- Escolhe a melhor sequência

Markov Chain

- Probabilidade do próximo rótulo só depende do anterior

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



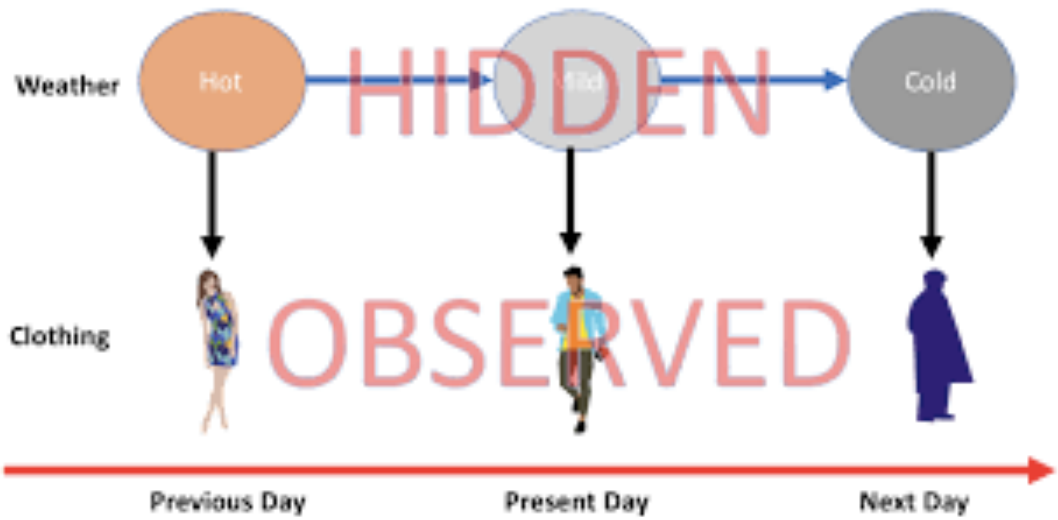


Markov Chain: Componentes

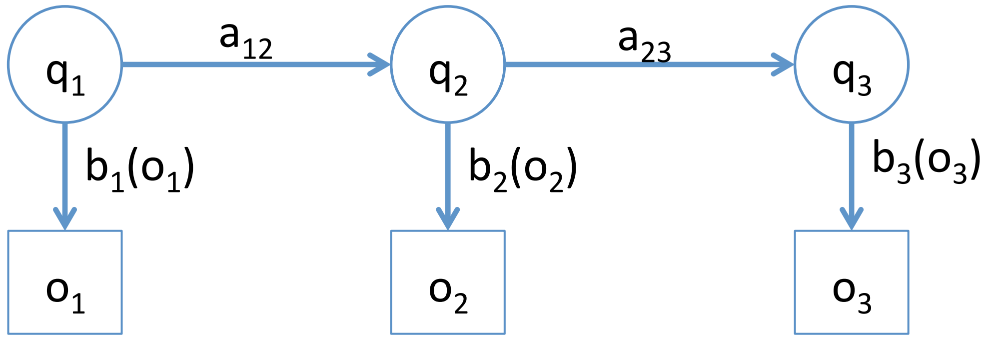
$Q = q_1 q_2 \dots q_N$	a set of N **states**
$A = a_{11} a_{12} \dots a_{N1} \dots a_{NN}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
$\pi = \pi_1, \pi_2, \dots, \pi_N$	an **initial 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, $\sum_{i=1}^{n} \pi_i = 1$

Hidden Markov Model

- Rótulos estão escondidos (hidden)
- Observa palavras
- Inferir rótulos (ex. POS) da sequência de palavras



Hidden Markov Model: Componentes



$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} \ldots a_{ij} \ldots a_{NN}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V = v_1, v_2, \ldots, v_V$
$B = b_i(o_t)$	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
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	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, $\sum_{i=1}^{n} \pi_i = 1$

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

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

HMM Tagger

- Matrix A: probabilidades de transição das tags

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

WSJ corpus

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

verbo modal

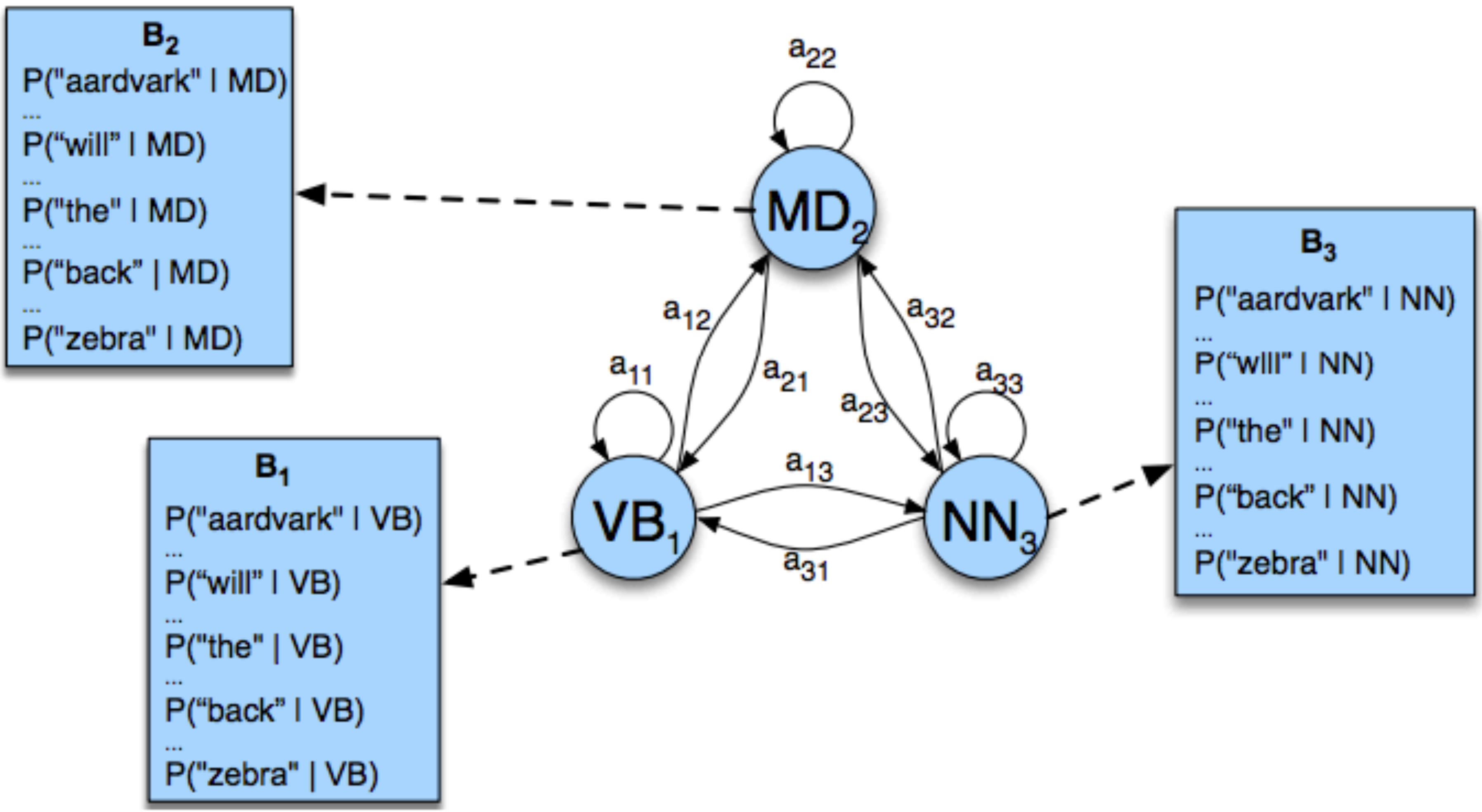
HMM Tagger

- Matrix B: probabilidades de uma palavra associada a uma tag

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

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

HMM Tagger: Exemplo



Decoding

- Dado um HMM como entrada: matrizes (A,B) e uma sequência de observações (palavras), encontrar a sequência de estados mais provável

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

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

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

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

Matriz B Matriz A

Viterbi Algorithm

- Input: Janet will back the bill
- Output: Janet/NNP will/MD back/VB the/DT bill/NN

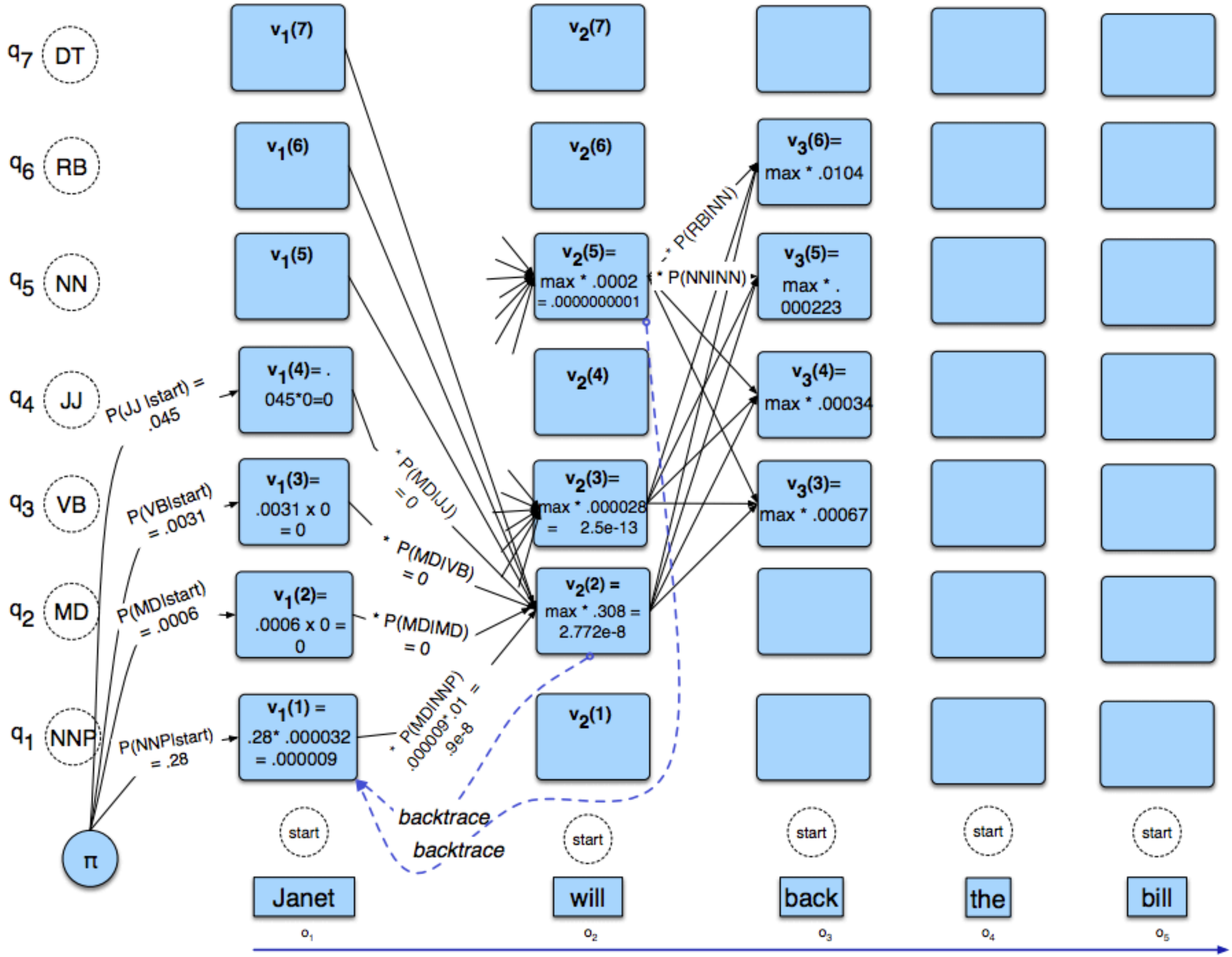
	NNP	MD	VB	JJ	NN	RB	DT
$<s>$	0.2767	0.0006	0.0031	0.0453	0.0449	0.0510	0.2026
NNP	0.3777	0.0110	0.0009	0.0084	0.0584	0.0090	0.0025
MD	0.0008	0.0002	0.7968	0.0005	0.0008	0.1698	0.0041
VB	0.0322	0.0005	0.0050	0.0837	0.0615	0.0514	0.2231
JJ	0.0366	0.0004	0.0001	0.0733	0.4509	0.0036	0.0036
NN	0.0096	0.0176	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

Matriz A (WSJ corpus)

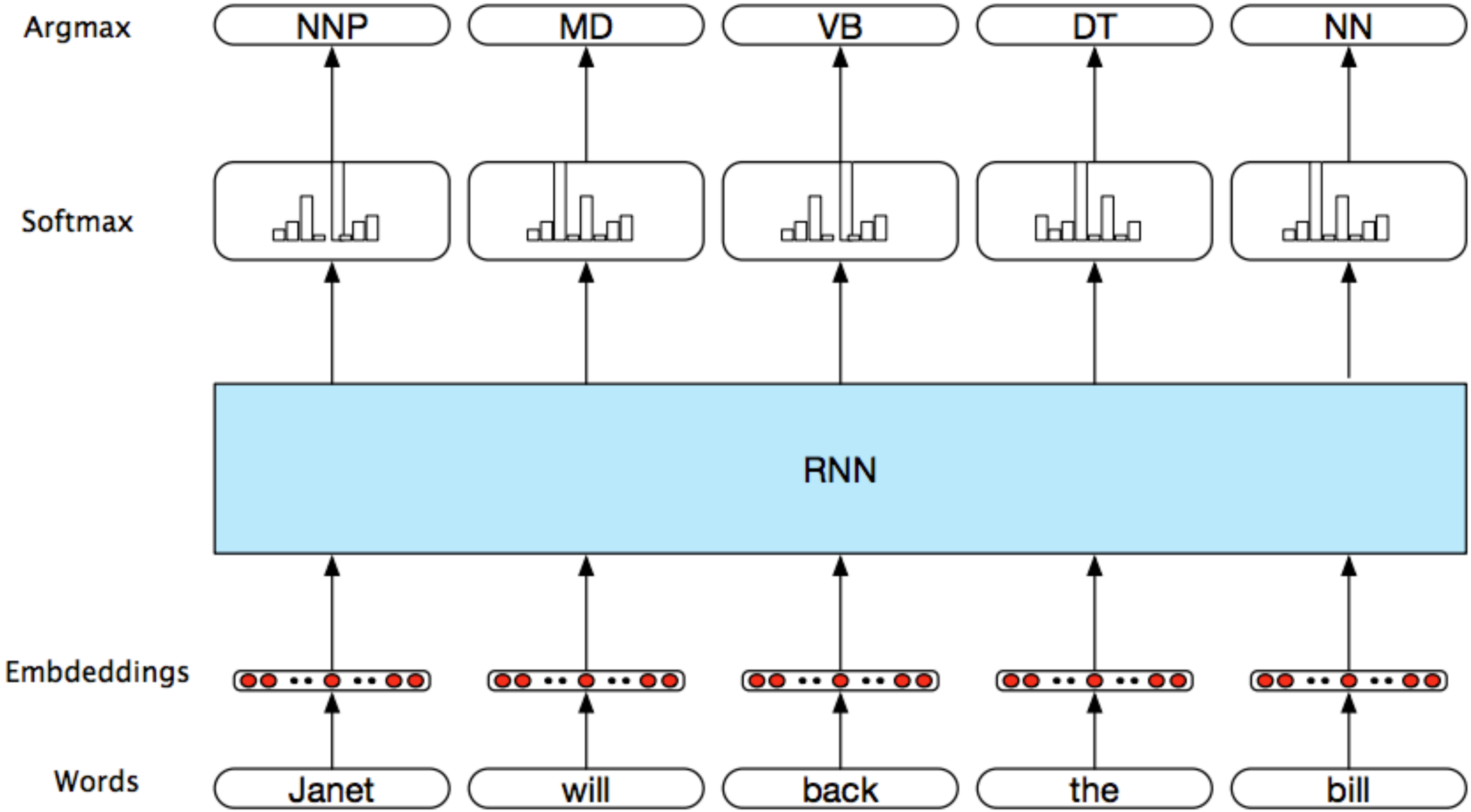
Viterbi Algorithm

	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

Matriz B (WSJ corpus)


q7 DT
q6 RB
q5 NN
q4 JJ
q3 VB
q2 MD
q1 NNP
π
P(JJ |start) = .045
P(VB|start) = .0031
P(MD|start) = .0006
P(NNP|start) = .28
v1(7)
v1(6)
v1(5)
v1(4)= . 045*0=0
v1(3)= .0031 x 0 = 0
v1(2)= .0006 x 0 = 0
v1(1) = .28* .000032 = .000009
* P(MD|JJ) = 0
* P(MD|VB) = 0
* P(MD|MD) = 0
* P(MD|NNP) .000009*.01 = .9e-8
v2(7)
v2(6)
v2(5)= max * .0002 = .0000000001
v2(4)
v2(3)= max * .000028 = 2.5e-13
v2(2) = max * .308 = 2.772e-8
v2(1)
* P(RB|NN)
* P(NN|NN)
v3(6)= max * .0104
v3(5)= max * . 000223
v3(4)= max * .00034
v3(3)= max * .00067
backtrace
backtrace
start
start
start
start
start
Janet
will
back
the
bill
o1
o2
o3
o4
o5

RNN para POS Tagging

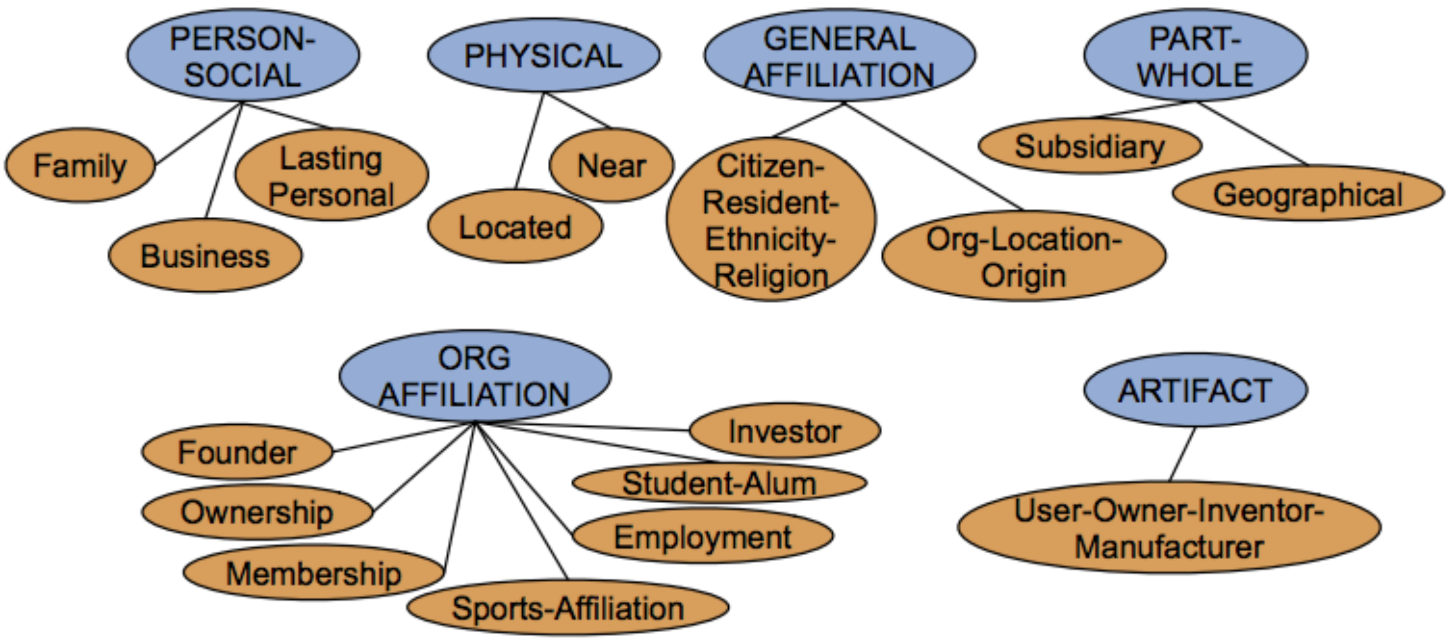


Relation Extraction

- Encontrar e classificar relações entre entidades

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**].

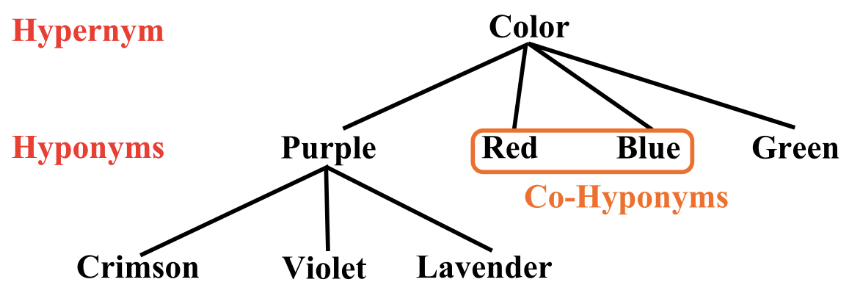
Exemplos de Relações



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**...

Extração Baseada em Padrões

- Hearst para extração de hyponyms



- Exemplo

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

- Padrão:

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

$$\downarrow$$

hyponym(Gelidium, red algae)

Extração Baseada em Padrões

- Hearst para extração de hyponyms

NP {, NP}* {,} (and|or) other NP_H	temples, treasuries, 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

Extração Baseada em ML

1. Definem-se as relações e entidades a serem extraídas
2. Anotam-se exemplos para treinamento

function FINDRELATIONS(*words*) **returns** *relations*

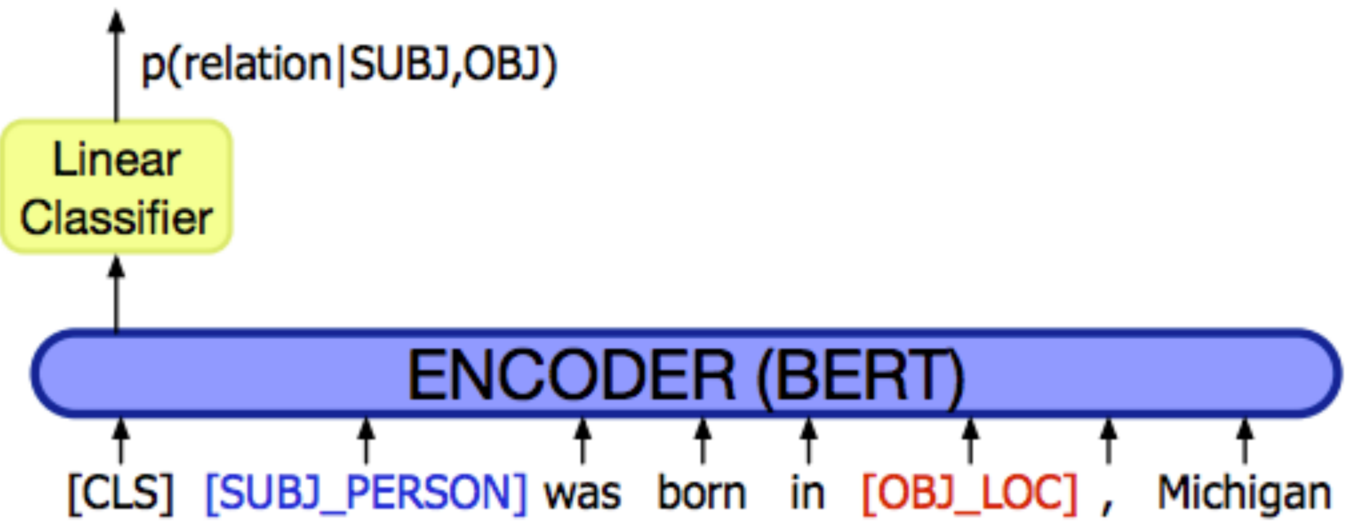
relations ← *nil*
entities ← FINDENTITIES(*words*)
forall entity pairs $\langle e1, e2 \rangle$ **in** *entities* **do**
  if RELATED?($e1, e2$)
    relations ← *relations*+CLASSIFYRELATION($e1, e2$)

Exemplos de Features

American Airlines, a unit of AMR, immediately matched the move, spokesman **Tim Wagner** said

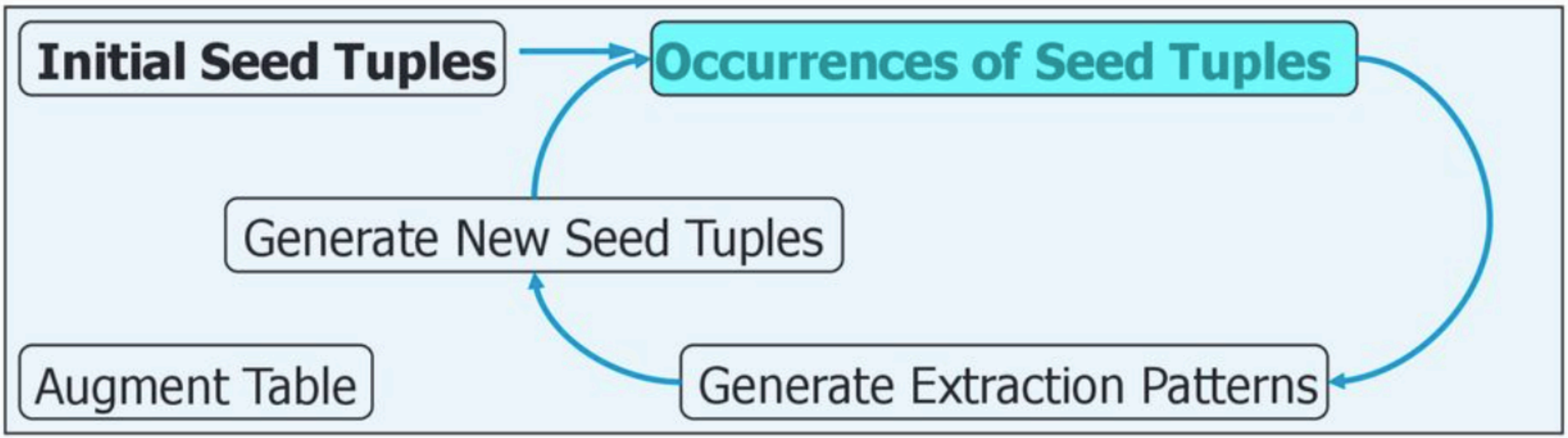
- Palavras ou bigramas ao redor
- Tipos das entidades
- Número de entidades entre as entidades candidatas
- Part-of-speech

Baseado em Rede Neural

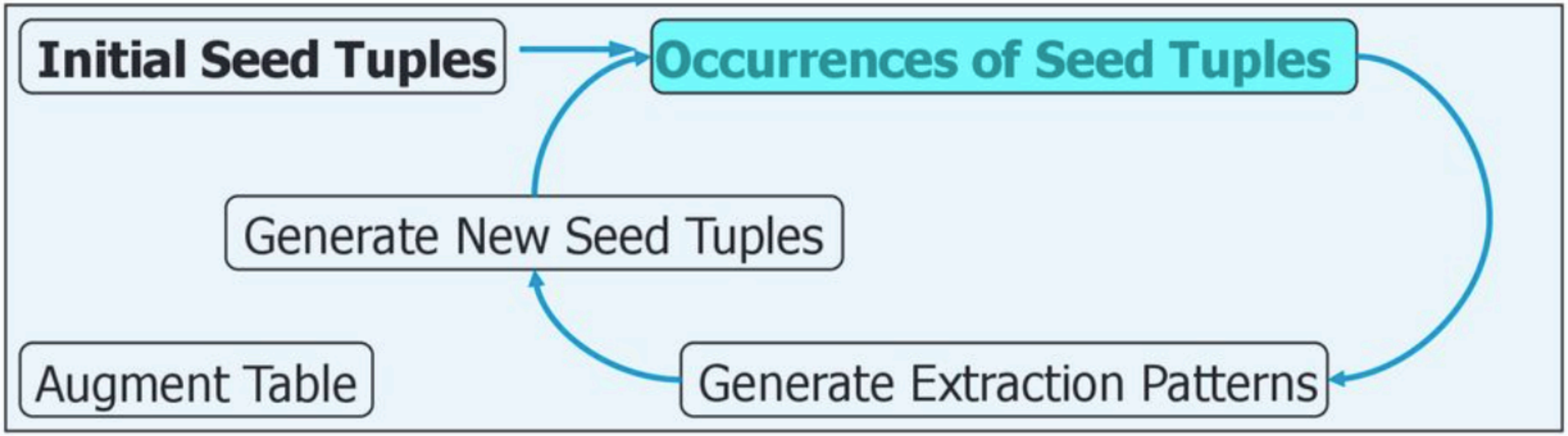


Bootstrapping

- Rótulos são difíceis de obter



Bootstrapping



Occurrences of seed tuples:

ORGANIZATION	*LOCATION*
MICROSOFT	REDMOND
IBM	ARMONK
BOEING	SEATTLE
INTEL	SANTA CLARA

Computer servers at Microsoft's headquarters in Redmond...

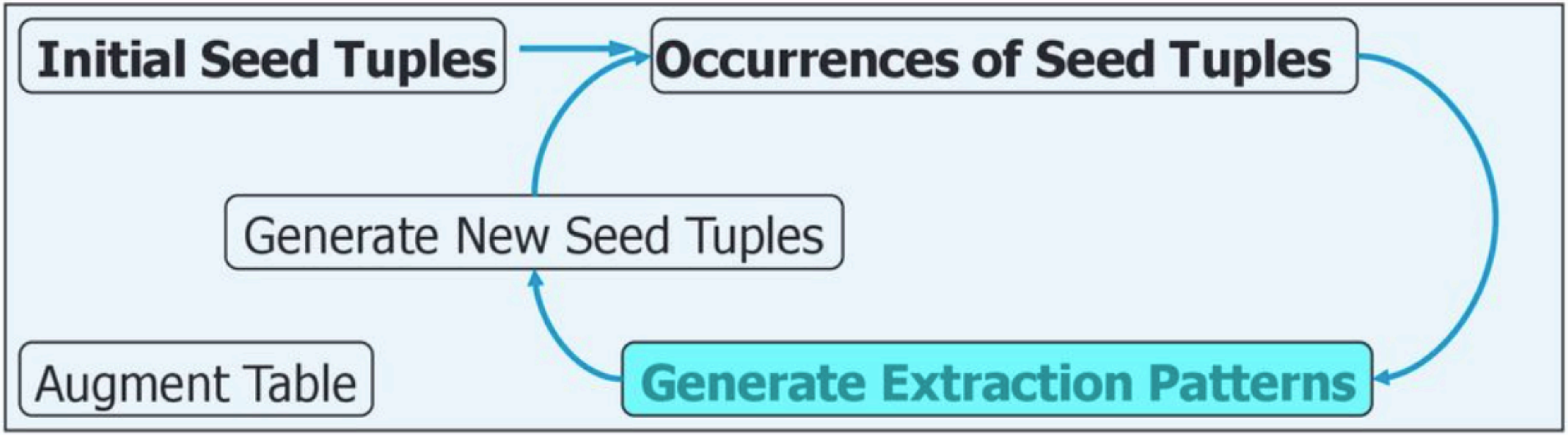
In mid-afternoon trading, share of Redmond-based Microsoft fell...

The Armonk-based IBM introduced a new line...

The combined company will operate from Boeing's headquarters in Seattle.

Intel, Santa Clara, cut prices of its Pentium processor.

Bootstrapping



Bootstrapping: Gerando Novas Sementes

- Uso dos padrões aprendidos

ORGANIZATION	*LOCATION*
AG EDWARDS	ST LUIS
157TH STREET	MANHATTAN
7TH LEVEL	RICHARDSON
3COM CORP	SANTA CLARA
3DO	REDWOOD CITY
JELLIES	APPLE
MACWEEK	SAN FRANCISCO