FACULDADE IBRATEC

Rotulação de Partes da Fala e Reconhecimento de Entidades

Prof. Marcelo Pita

Pós em Inteligência Artificial

Processamento de Linguagem Natural

Rotulação de partes da fala

Do inglês ***Part-of-Speech (POS) Tagging***.

Partes da fala são classes de palavras, categorias morfo-sintáticas ou categorias gramaticais.

- Exemplos: substantivos, verbos, adjetivos, advérbios, pronomes, etc.

A **Rotulação (ou etiquetação) de partes da fala** consiste na *atribuição de partes da fala a palavras*.

Partes da Fala

Substantivo

Adjetivo

Verbo

Advérbio

Preposição

Pronome

...

Rotuladores automáticos

Abordagens principais:

- Baseadas em regras
- Probabilísticas

Modelos **probabilísticos**:

1. Unigramas
2. Bigramas
3. Trigramas
4. Modelos ocultos de Markov

Rotuladores automáticos

Modelo probabilístico baseado em unigramas

Seleciona classe gramatical mais frequente para um token em uma base de treinamento.

$$P(t_i|w) = \frac{c(w, t_i)}{c(w, t_1) + ... + c(w, t_k)}$$

Rotuladores automáticos

Modelo probabilístico baseado em bigramas

Considera a frequência de ocorrência de bigramas das classes gramaticais de um token e seu antecessor. Selecionar classe gramatical que maximize a seguinte equação:

$$\prod_{i=1}^{n} P(t_i|t_{i-1})P(w_i|t_i)$$

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

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

Rotuladores automáticos

Modelo probabilístico baseado em trigramas

Similar a de bigrama, porém considera também a classe gramatical do antecessor do antecessor. Maximizar a equação:

$$\prod_{i=1}^{n} P(t_i|t_{i-2}, t_{i-1})P(w_i|t_i)$$

Rotuladores automáticos

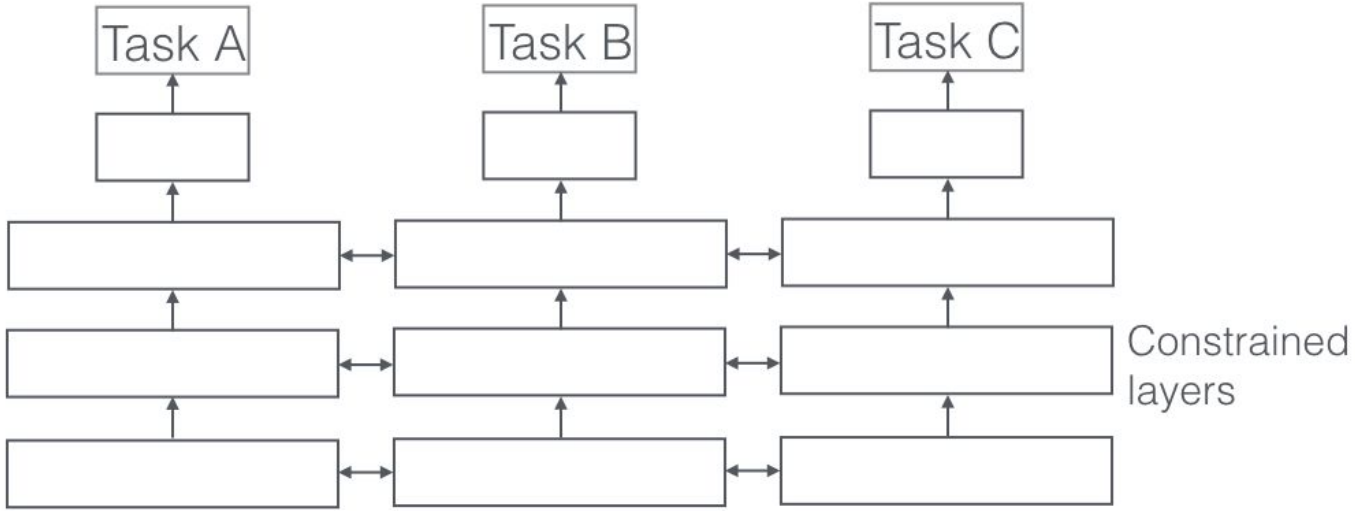
Modelo oculto de Markov (*Hidden Markov Model - HMM*)

O objetivo é encontrar a sequência de partes da fala com a maior probabilidade, dada uma sequência de tokens.

$$T = \arg\max_{\hat{T}} P(\hat{T}|W) = \arg\max_{\hat{T}} \frac{P(W|\hat{T})P(\hat{T})}{P(W)} \propto \arg\max_{\hat{T}} P(W|\hat{T})P(\hat{T})$$

Rotuladores automáticos

Modelo neural baseado em redes neurais convolutivas



Rotuladores automáticos

<Prática>

Reconhecimento de Entidades

Do inglês ***Named Entity Recognition (NER)***.

Entidades são categorias semânticas de palavras.

- Exemplos: datas, lugares, pessoas, organizações, etc.

O **reconhecimento de entidades nomeadas** consiste na *identificação e classificação automática de partes do texto como entidades*.

Entidades Nomeadas

Pessoas

Lugares

Datas

Horas

Organizações

Valores (dinheiro)

...

Reconhecimento de Entidades

Exemplo:

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**].

Reconhecimento de Entidades

Métodos podem apresentar erros em situações de ambiguidade.

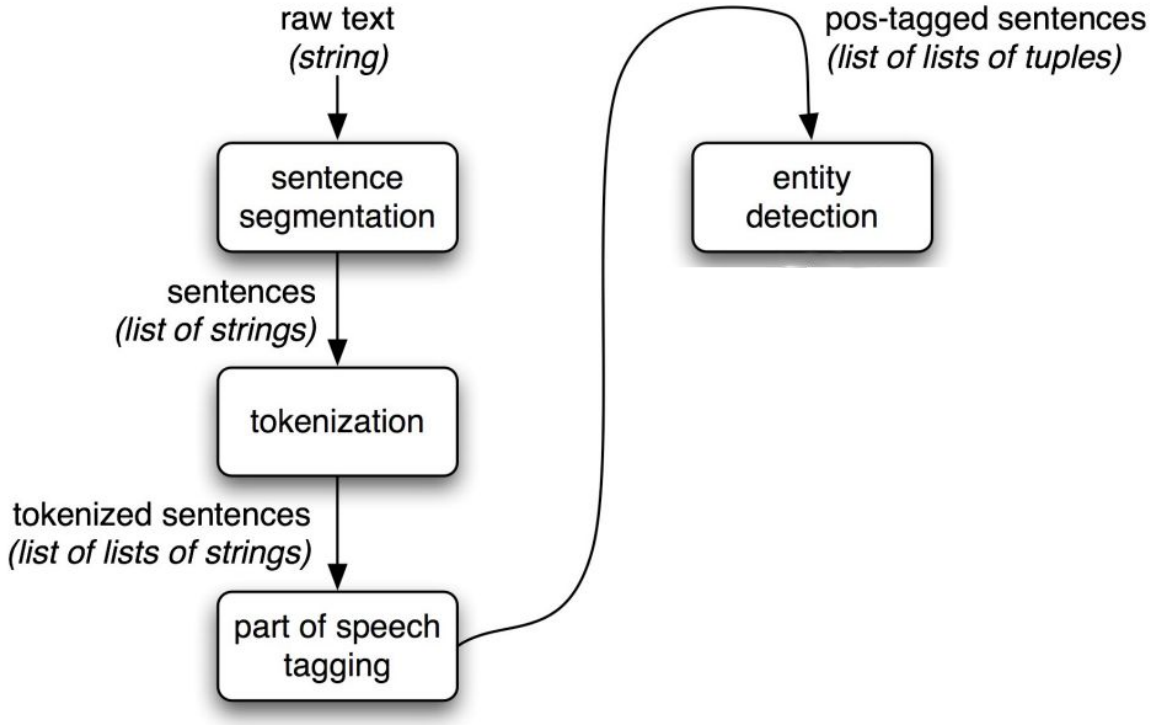
Juscelino Kubitschek pode ser:

- *Pessoa*
- *Organização* (Escola, Governo)
- *Lugar* (Avenida)

Desambiguação envolve análise de contexto.

Reconhecimento de Entidades

Fluxo padrão para reconhecimento de entidades.



Reconhecimento de Entidades

Principais algoritmos:

- Baseado em *features*
- Redes neurais
- Baseados em regras

Avaliação:

- Métricas padrão de classificadores: F1, recall, precision

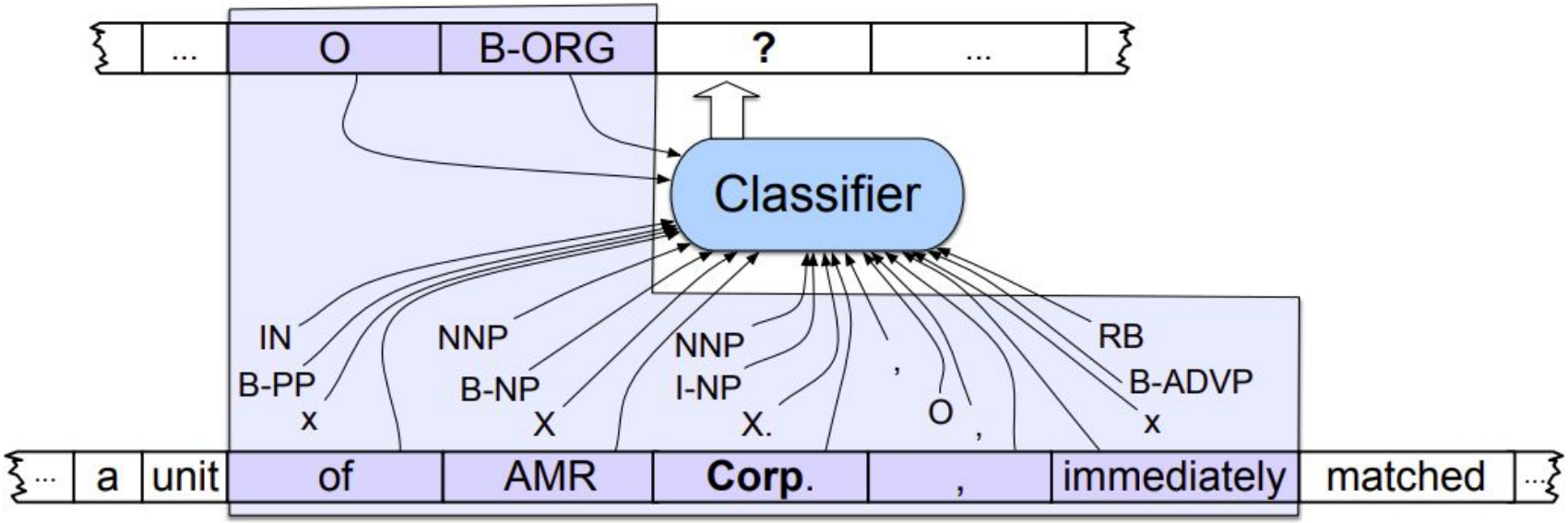
Reconhecimento de Entidades

Representação de entidades nomeadas (rotulação IOB)

Words	IOB Label	IO Label
American	B-ORG	I-ORG
Airlines	I-ORG	I-ORG
,	O	O
a	O	O
unit	O	O
of	O	O
AMR	B-ORG	I-ORG
Corp.	I-ORG	I-ORG
,	O	O
immediately	O	O
matched	O	O
the	O	O
move	O	O
,	O	O
spokesman	O	O
Tim	B-PER	I-PER
Wagner	I-PER	I-PER
said	O	O
.	O	O

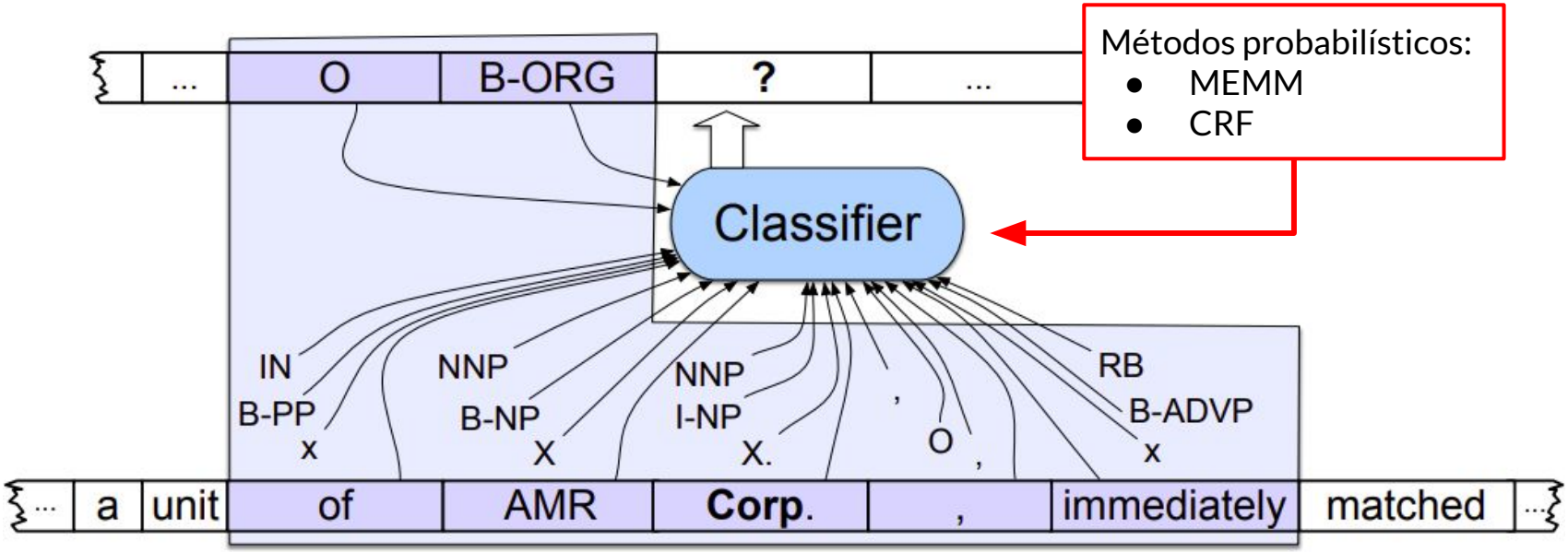
Reconhecimento de Entidades

Algoritmo baseado em *features*.



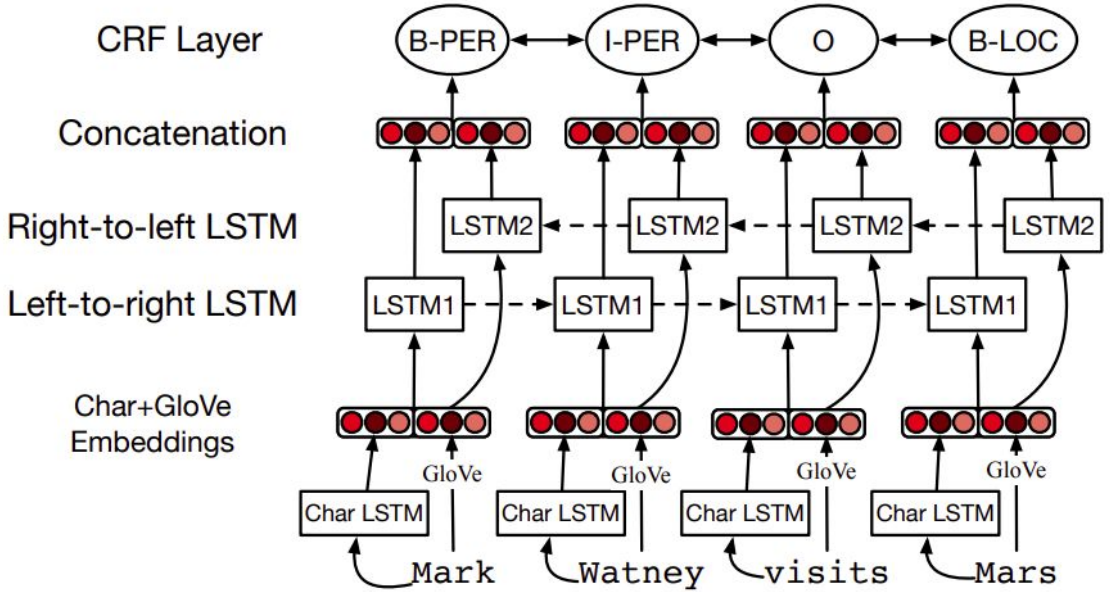
Reconhecimento de Entidades

Algoritmo baseado em *features*.



Reconhecimento de Entidades

Algoritmo baseado em redes neurais (bi-LSTM).



Reconhecimento de Entidades

Abordagem prática *muito usada no mercado* que mistura o casamento de palavras com regras e modelos probabilísticos para classificação de sequência.

- Aplicar regras de alta precisão (baixos FP, mas alto FN)
- Buscar casamento de regras com substrings do nomes já detectados
- Buscar palavras em listas de nomes específicas de domínio
- Aplicar classificação de sequência (MEMM, CRF) incluindo as tags descobertas nas fases anteriores

Reconhecimento de Entidades

<Prática>