Linear-Chain Conditional Random Fields (CRF)

- Intuição, usos e aprendizados

Fábio Capuano de Souza
fabiocapsouza@gmail.com

CRF: motivação

- Problemas que precisam prever uma **sequência de saídas** para uma sequência de entradas.
- Exemplos:
  - Segmentação: classificar cada pixel de uma imagem
    - Há uma relação grande entre a classe de pixels vizinhos
  - NLP: *Sequence Labeling*
    - Classificar cada token de um texto em um vocabulário de Tags
    - Part of Speech (POS), NER (Named Entity Recognition), ...

Paulo	comprou	300	ações	da	empresa	ACME	Corp	.	em	2006	.
B-PER	O	B-VAL	O	O	O	B-ORG	I-ORG	I-ORG	O	B-TEMPO	O

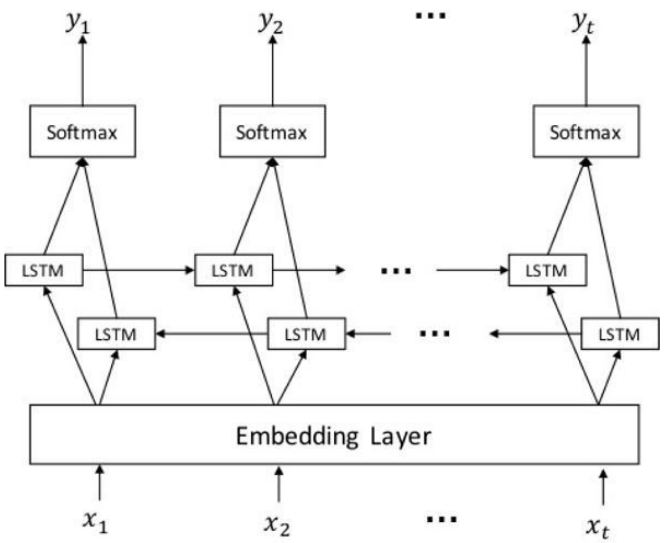
NER: modelamento

- *Sequence tagging*: tarefa a **nível de token** (*token-level*) onde deve-se classificar cada **token** da sequência de entrada
  - Identificação e classificação **simultâneas** das entidades
- Esquema de tagging:
  - Vocabulário de tags, com tags distintas para cada classe
  - IOB2: B (begin), I (in), O (out)
    - O, B-PER, I-PER, B-ORG, I-ORG, etc

Paulo	comprou	300	ações	da	empresa	ACME	Corp	.	em	2006	.
B-PER	O	B-VAL	O	O	O	B-ORG	I-ORG	I-ORG	O	B-TEMPO	O

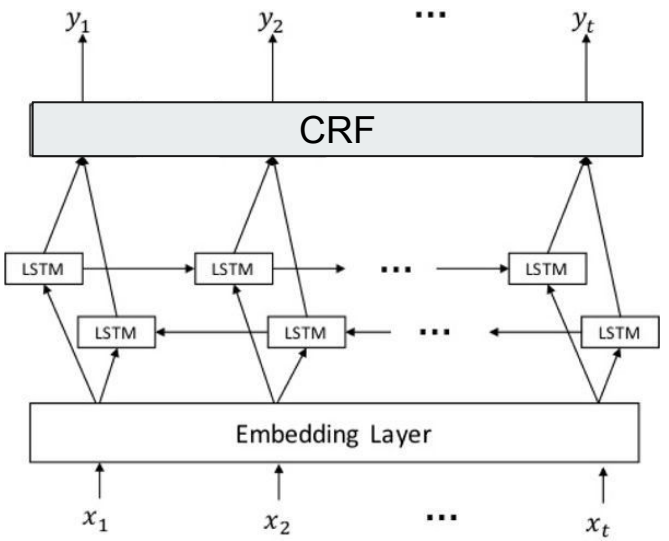
CRF: motivação

- As saídas de redes neurais recorrentes, como LSTM, são independentes entre si
  - $P(y_i \mid x_1, \ldots, x_n)$
- Porém, as labels nem sempre são de fato independentes
- Sequence Labeling impõe **restrições rígidas** nas labels
  - Tags **I-** nunca podem suceder tag **O**
  - Tag **I-PER** não pode suceder **B-ORG** ou **I-ORG**



CRF

- O CRF permite incluir dependência entre labels
  - Linear-Chain: “$P(y_i \mid \mathbf{X}, y_{i-1})$”
- O CRF é um modelo para prever **a sequência de labels mais provável** para uma sequência de entradas
  - $P(\mathbf{Y} \mid \mathbf{X})$



CRF: Transições

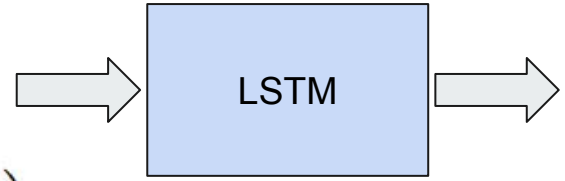
- Parâmetros do CRF
  - Matriz de **Transição entre classes + Transições inicial** e **final**
  - **Aij : custo de transicionar da classe i para a classe j**
  - K classes: $K^2 + 2K$ parâmetros

A =

C(cont. em A)	C(ir B p/ A)	C(ir C p/ A)	C(começar c/ A)	C(acabar em A)
C(ir A p/ B)	C(cont. em B)	C(ir C p/ B)	C(começar c/ B)	C(acabar em B)
C(ir A p/ C)	C(ir B p/ C)	C(cont. em C)	C(começar c/ C)	C(acabar em C)

CRF: função de custo

$$\begin{cases} \mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) \\ \mathbf{y} = (y_1, y_2, \ldots, y_n) \end{cases}$$



P : matriz de scores de tamanho n x K "logits"

Dados **P** e **y**, o score de uma *sequência de predições* **y** é dado por: (y0: *start* e yn: *end*)

$$s(\mathbf{X}, \mathbf{y}) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i}$$

CRF: Exemplo de score

$$s(\mathbf{X}, \mathbf{y}) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i}$$

y = (A, B, B); n = 3; K=2

$$\mathbf{P} = \begin{bmatrix} P_{1,A} & P_{2,A} \\ P_{2,A} & \\ P_{1,B} & P_{2,B} \\ P_{2,B} & \end{bmatrix}$$

$$s(\mathbf{X}, \mathbf{y}) = A_{start,\,A} + P_{1,A} + A_{A,\,B} + P_{2,B} + A_{B,B} + P_{3,B} + A_{B,\,end}$$

CRF: Função de custo

- n = 3; K=2
- K^n possíveis sequências **y**
- $2^3 = 8$ scores

$$s(\mathbf{X}, (A, A, A)) = A_{start, A} + P_{1,A} + A_{A, A} + P_{2,A} + A_{A,A} + P_{3,A} + A_{A, end}$$

$$s(\mathbf{X}, (A, B, B)) = A_{start, A} + P_{1,A} + A_{A, B} + P_{2,B} + A_{B,B} + P_{3,B} + A_{B, end}$$

(...)

$$s(\mathbf{X}, (B, B, B)) = A_{start, B} + P_{1,B} + A_{B, B} + P_{2,B} + A_{B,B} + P_{3,B} + A_{B, end}$$

Probabilidade de uma sequência:
Softmax entre todas as possíveis sequências $\mathbf{Y}_X$

$$p(\mathbf{y}|\mathbf{X}) = \frac{e^{s(\mathbf{X},\mathbf{y})}}{\sum_{\tilde{\mathbf{y}} \in \mathbf{Y}_\mathbf{X}} e^{s(\mathbf{X},\tilde{\mathbf{y}})}}.$$

CRF: Função de custo

Custo: maximizar o log da probabilidade da sequência correta

Probabilidade da **sequência correta y**:
Softmax entre todas as possíveis sequências $\mathbf{Y_X}$

$$p(\mathbf{y}|\mathbf{X}) = \frac{e^{s(\mathbf{X},\mathbf{y})}}{\sum_{\widetilde{\mathbf{y}}\in\mathbf{Y_X}} e^{s(\mathbf{X},\widetilde{\mathbf{y}})}}.$$

$$\begin{aligned}\log(p(\mathbf{y}|\mathbf{X})) &= s(\mathbf{X},\mathbf{y}) - \log\left(\sum_{\widetilde{\mathbf{y}}\in\mathbf{Y_X}} e^{s(\mathbf{X},\widetilde{\mathbf{y}})}\right) \\ &= s(\mathbf{X},\mathbf{y}) - \underset{\widetilde{\mathbf{y}}\in\mathbf{Y_X}}{\text{logadd}}\, s(\mathbf{X},\widetilde{\mathbf{y}}), \qquad (1)\end{aligned}$$

CRF: Treinamento e Inferência

- Treinamento vai aprender os custos das transições **A**
- Após o treinamento, a predição para um exemplo precisa encontrar a sequência de maior score

$$\mathbf{y}^* = \underset{\tilde{\mathbf{y}} \in \mathbf{Y_X}}{\operatorname{argmax}}\, s(\mathbf{X}, \tilde{\mathbf{y}}).$$

- Cálculo do score de todas as sequências é intratável se feito de forma “inocente”
  - Algoritmos eficientes com programação dinâmica
  - Cálculo de todos os scores: *Forward-backward algorithm*
  - Predição: *Viterbi Decoding*

PyTorch: biblioteca

- pytorch-crf: https://pytorch-crf.readthedocs.io/en/stable/
- Camada CRF:
  - **`forward`**(*emissions*, *tags*, *mask=None*, *reduction='sum'*)
    - Usado somente durante o treinamento
    - Retorna a log-probability da sequência correta -> **custo**
  - **`decode`**(**emissions, mask=None**)
    - Inferência com Viterbi Decoding
    - Retorna a sequência de tags mais provável

Impacto do CRF em NER

Table 3: Results of NER task (Precision, Recall and micro F1-score) on the test set (MiniHAREM). Best results in bold. Reported values are the average of multiple runs with different random seeds. (*): primary metric.

Architecture	Total scenario			Selective scenario		
	Prec.	Rec.	F1 (*)	Prec.	Rec.	F1 (*)
CharWNN [22]	67.2	63.7	65.4	74.0	68.7	71.2
LSTM-CRF [3]	72.8	68.0	70.3	78.3	74.4	76.3
BiLSTM-CRF+FlairBBP [24]	74.9	74.4	74.6	83.4	81.2	82.3
mBERT	71.6	72.7	72.2	77.0	78.8	77.9
mBERT + CRF	74.1	72.2	73.1	80.1	78.3	79.2
BERTimbau Base	76.8	77.1	77.2	81.9	**82.7**	82.2
BERTimbau Base + CRF	78.5	76.8	77.6	84.6	81.6	83.1
BERTimbau Large	77.9	**78.0**	77.9	81.3	82.2	81.7
BERTimbau Large + CRF	**79.6**	77.4	**78.5**	**84.9**	82.5	**83.7**

Colab com implementação de exemplo

Toy example:

https://colab.research.google.com/drive/1xtS5Wts8JEQDHf7VGs-9VQJFknLEvSeF