Relatório Semestral do Projeto: Séries Temporais para Dados do Judiciário

Autor: Gabriel Dalforno Silvestre
Orientador: André C. P. L. F. de Carvalho

Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

Palavras-Chave: Séries Temporais, Predição, Aprendizado de Máquina, Mineração de Dados.

1 Introdução

As séries temporais que a príncipio seriam utilizadas como objeto de estudo deste projeto ainda não foram disponibilizadas. Enquanto isso, o estudo foi redirecionado para a tarefa de reconhecimento de entidades nomeadas (REN), no âmbito de processamento de linguagem natural. O trabalho se resumi a buscar todas as referências a documentos legislativos em Projetos de Lei (PL's) e Solicitações de Trabalho (ST's) escritas por membros da Câmara dos Deputados. Este relatório esta dividido da seguinte forma: a Seção 2 introduz o conceito de REN junto as técnicas mais utilizadas na literatura, a Seção 3 contém os primeiros resultados de aplicação do REN às PL's e ST's e por fim, a Seção 4 ilustra o estado atual da tarefa.

2 Reconhecimento de Entidades Nomeadas

O objetivo do REN é atribuir a cada *token* de uma sentença um respectivo rótulo de acordo com algum critério pré-estabelecido de forma automática [1].

Considere a tarefa de mapear cada palavra de uma sentença na sua respectiva classe gramatical, por exemplo, se tomarmos a frase: "*O cachorro está latindo para o gato*", gostaríamos de obter a seguinte resposta:

1. $O \mapsto$ *artigo*;
2. *cachorro* $\mapsto$ *substantivo*;
3. *está* $\mapsto$ *verbo*;
4. *latindo* $\mapsto$ *verbo*;
5. *para* $\mapsto$ *preposição*;
6. *o* $\mapsto$ *artigo*;
7. *gato* $\mapsto$ *substantivo*;

Repetir esse processo de forma manual para um conjunto grande de sentenças é inviável do ponto de vista prático, porém com algoritmos especializados neste tipo de tarefa que adquirem conhecimento de forma automática a partir de um conjunto de dados rotulado, é possível atingir resultados muito próximos à rotulagem manual.

As duas abordagens mais comuns para modelagem de conjunto de dados em REN são as generativas e as discriminativas.

2.1 Abordagem Generativa

Seja $\mathcal{V}$ um vocabulário e $\mathcal{K}$ um conjunto finito de *tags* (rótulos associados cada palavra numa sentença). Defina S como o conjunto de todas os pares sentença/*tag* como em 1

$$S = \{(x_1, \ldots, x_n, y_1, \ldots, y_n) : x_i \in \mathcal{V}, y_i \in \mathcal{K}, \forall i = 1, \ldots, n\} \tag{1}$$

Um modelo generativo é uma função p satisfazendo as seguintes propriedades:

- $p(x_1, \ldots, x_n, y_1, \ldots, y_n) > 0, \forall (x_1, \ldots, x_n, y_1, \ldots, y_n) \in \mathcal{S}$
- $\sum_{(x_1,\ldots,x_n,y_1,\ldots,y_n)\in\mathcal{S}} p(x_1, \ldots, x_n, y_1, \ldots, y_n) = 1$

Nessas condições, definimos o *tagger* f como em 2

$$f(x_1, \ldots, x_n) = \arg \max_{y_1, \ldots, y_n} p(x_1, \ldots, x_n, y_1, \ldots, y_n) \tag{2}$$

Um dos exemplos mais comuns deste tipo de modelo são os Modelos Ocultos de Markov (MOM). Este considera uma suposição de Markov de ordem 2, neste caso tem-se que p se reduz a Equação 3.

$$p(x_1, \ldots, x_n, y_1, \ldots, y_n) = \prod_{i=1}^{n+1} q(y_i | y_{i-1}, y_{i-2}) \prod_{i=1}^{n} e(x_i | y_i) \tag{3}$$

Denotando por $c(u, v, s)$ o número de vezes que a sequência (u, v, s) aparece no conjunto de treinamento, $c(u, v)$ o número de aparições do par (u, v), $c(s \rightsquigarrow x)$ o número de vezes que a *tag* s aparece pareada com o token x e $c(x)$ o número de aparições do token x define-se q e e como em 4

$$\begin{aligned} q(s|u, v) &= \frac{c(u, v, s)}{c(u, v)} \\ e(x|s) &= \frac{c(s \rightsquigarrow x)}{c(x)} \end{aligned} \tag{4}$$

2.2 Abordagem discriminativa

Considere um conjunto de possível *inputs* $\mathcal{X}$, de possíveis rótulos $\mathcal{Y}$, uma função $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ que mapeia um par (x, y) num vetor de atributos e um parâmetro v. Um model discriminativo assume a forma dada por 5

$$p(y|x; v) = \frac{\exp(\langle v, g(x, y) \rangle)}{\sum_{y' \in \mathcal{Y}} \exp(\langle v, g(x, y') \rangle)} \tag{5}$$

Neste caso, o processo de treinamento se reduz a estimar o parâmetro v^* que maximiza a função log verossimilhança dado um conjunto de dados $\{(x_i, y_i) : x_i \in \mathcal{X}, y_i \in \mathcal{Y}, \forall i = 1, \ldots, n\}$ como na Equação 6.

$$v^* = \arg \max_{v \in \mathbb{R}^d} \sum_{i=1}^{n} \log(p(y_i | x_i; v)) \tag{6}$$

Tendo v^* em mãos, o *tagger* f é definido como mostra a Equação 7

$$f(x) = \arg \max_{y' \in \mathcal{Y}} p(y' | x; v^*) \tag{7}$$

Segundo a literatura, um dos modelos discriminativos que obteve mais sucesso em resultados experimentais nas tarefas de REN é o *Conditional Random Fields (CRF).*

3 Primeiros Resultados

Diferentemente do exemplo dado no ínicio da Seção 2, estamos interessados na busca de referências à documentos legistativos em textos legais. Inicialmente foram consideradas somente duas possíveis *tags* para cada *token* nos conjuntos de dados disponibilizados pela câmara dos deputados, a *tag* "DOCUMENTO" e a *tag* "O". A primeira refere-se à documentos referenciados por membros da câmara e a segunda à outras palavras.

Foram considerados dois conjuntos de dados para análise inicial, o das PL's e o das ST's, após processamento, limpeza e divisão destes conjuntos de dados, obtemos os seguintes números ilustrados na Tabela 1.

Propriedade	**PL's**	**ST's**
# Total de sentenças	4032	680
# Sentenças no conjunto de treinamento	2822	475
# Sentenças no conjunto de teste	1210	205
# Tokens únicos no conjunto de treinamento	11017	4256
# Tokens rotulados como "O" no conjunto de treinamento	105578	15716
# Tokens rotulados como "DOCUMENTO" no conjunto de treinamento	3142	1101

Table 1: Informações gerais à respeito das PL's e das ST's.

As duas abordagens: generativa e discriminativa, tendo como representantes o MOM e o CRF, respectivamente, foram comparadas nos dois conjuntos de dados. Os modelos tiveram seus parâmetros estimados utilizando os conjuntos de treinamento e suas capacidades preditivas avaliadas nos conjuntos de teste.

As tabelas 2 e 3 mostram as matrizes de confusão dos dois modelos após serem validados nos conjuntos de teste das PL's e das ST's. Para simplificar a notação estamos denotando a *tag* "DOCUMENTO" simplesmente por D e o símbolo $\hat{D}$ indica as predições dos modelos para a *tag* D, a interpretação é a mesma para $\hat{O}$.

Nota-se que o CRF foi superior ao MOM nos dois conjuntos de dados, obtendo um certo balanço entre o número de falso positivos e falso negativos, o que mostra que o modelo é robusto com respeito ao desbalanceamento das

	D	O
$\hat{D}$	291	348
$\hat{O}$	183	21874

	D	O
$\hat{D}$	9	162
$\hat{O}$	3	3927

Table 2: Matrizes de Confusão do MOM, à esquerda os resultados nas PL's e à direita os resultados nas ST's.

	D	O
$\hat{D}$	500	139
$\hat{O}$	120	21937

	D	O
$\hat{D}$	131	40
$\hat{O}$	24	3906

Table 3: Matrizes de Confusão do CRF, à esquerda os resultados nas PL's e à direita os resultados nas ST's.

classes, muito comum em tarefas de REN. Por outro lado, o MOM sofreu significativamente mais com aquele, principalmente nas ST's onde quase se reduziu à um modelo constante $f(x) =$ "O". A precisão balanceada, calculada de acordo com a Equação 8, reforça o que foi inferido a partir das matrizes de confusão como pode ser visto na tabela 4.

$$\frac{1}{2}\Big(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\Big) \tag{8}$$

Modelo	**PL's**	**ST's**
MOM	72.36%	52.59%
CRF	88.51%	88.00%

Table 4: Precisão balanceada do MOM e do CRF nos conjuntos de dados das PL's e das ST's.

4 Próximos Passos

O próximo passo é refinar as *tags* nestes conjuntos de dados com o objetivo de abranger mais entidades como referências à pessoas, eventos, cargos, projetos de lei, eventos, etc. Uma vez terminada a fase de rotulagem, pode-se partir para a análise do impacto dessas mudanças na performance dos modelos de aprendizado de máquina considerados e dirigir os esforços para tornar as técnicas com maiores chances de sucesso mais robustas para serem utilizadas à nível de produção.

References

[1] Michael Collins. Natural language processing, lecture notes. http://www.cs.columbia.edu/ mcollins/, 2021, Acessed: 2021-05-10.