Laboratório de Inteligência Artificial Aplicada da 3.ª Região - LIAA-3R

SINARA – Resolvendo problemas de linguagem natural em textos jurídicos através da extração de dispositivos normativos e precedentes.

São Paulo, fevereiro de 2020

Laboratório de Inteligência Artificial Aplicada da 3.ª Região - LIAA-3R

SINARA – Resolvendo problemas de linguagem natural em textos jurídicos através da extração de dispositivos normativos e precedentes.

Projeto de pesquisa apresentado ao Centro de
Inteligência Artificial aplicada
ao sistema Processo Judicial Eletrônico – PJe.

São Paulo, fevereiro de 2020

Sumário

1.Identificação do projeto.....04

2.Introdução.....08

3.Problema de Pesquisa.....09

4.Justificativa.....09

5.Objetivos.....10

6.Metodologia.....10

7.Cronograma.....11

8. Referências preliminares 12

1. IDENTIFICAÇÃO DO PROJETO

Título: SINARA – Resolvendo problemas de linguagem natural em textos jurídicos através da extração de dispositivos normativos e precedentes.

Grupo de pesquisa:

NOME	CPF	LOTAÇÃO	PERFIL
AKI ANDO KOJIMA	799.606.774-00	SJSP	Validação Ética e Jurídica de Modelos, Programadora Pré-Processamento e Data Augmentation e Anotadora
AUGUSTO CÉSAR DE CASTRO	118.171.738-80	TRF3R	Cientista de IA
CAIO MOYSÉS DE LIMA	163.371.628-70	SJSP	Coordenador, Programador Pré-Processamento e Data Augmentation e Curadoria
CLÁUDIO ROBERTO NÓBREGA MARTINS	175.944.848-60	TRF3R	Documentação e Apoio Operacional
CLAYTON MATOS DA SILVA	283.766.348-44	SJSP	Anotador
DANIEL CARNEIRO SILAMI	087.095.767-89	TRF3R	Anotador
DAVID PANESSA BACCELLI	299.071.118-05	TRF3R	Curadoria e Apoio Operacional
DAYANA ROSA DOS SANTOS	225.783.118-70	TRF3R	Anotadora

ELISA EMIKO TANAKA DELLI PAOLI	338.166.478-63	SJSP	Curadoria
ELIVAN DE MELO LIMA	061.146.054-89	SJSP	Anotador
FÁBIO AKAHOSHI COLLADO	224.366.648-03	TRF3R	Gestor Técnico, Cientista de IA e Curadoria
FREDERICO AGRÍCOLA BATISTA DA SILVA	083.763.416-42	TRF3R	Anotador
GABRIELA HARA	247.902.518-70	TRF3R	Anotadora
GISELE MOLINARI FESSORE	010.501.798-11	SJSP	Anotadora
GIULIA YURIKO TANAKA	053.779.769-67	SJSP	Programadora Pré-Processamento e Data Augmentation, Curadoria, Documentação e Apoio Operacional
JOÃO PAULO TIVERON	344.780.228-60	SJSP	Cientista de Dados e Programador Pré-Processamento e Data Augmentation
LUCIANA ORTIZ TAVARES COSTA ZANONI	141.291.568-61	SJSP	Validação Ética e Jurídica dos Modelos e Curadoria
LUIZ GUILHERME MARTINS	097.909.418-62	SJSP	Curadoria
MAÍRA ZAU SERPA SPINA	291.136.408-28	TRF3R	Validação Ética e Jurídica dos

D'EVA			Modelos, Documentação e Apoio Operacional
MARIA ALICE LEIS OLIVARES	266.646.688-40	TRF3R	Anotadora, Documentação e Apoio Operacional
MARIANE AKEMI NORISSADA	029.463.836-95	TRF3R	Anotadora
MATHEUS HENRIQUE DE PAIVA CARVALHO	327.190.028-01	TRF3R	Validação Ética e Jurídica dos Modelos e Curadoria
NATÁLIA TAVARES AMATO	173.186.218-07	SJSP	Anotadora
PAULO CEZAR NEVES JUNIOR	173.116.558-70	SJSP	Validação Ética e Jurídica dos Modelos e Curadoria
PEDRO HENRIQUE LOPES GUERRA	308.093.368-04	SJSP	Cientista de Dados
RAFAEL ARRUTI ARAGÃO VIEIRA	035.515.335-12	TRF3R	Cientista de Dados
RAQUEL KIRCHHEIM	722.258.900-72	TRF3R	Anotadora
REGINALDO MITSUO IWAMOTO	014.032.318-00	SJSP	Anotador
RENATA DE SOUZA PLENS	358.848.588-09	SJSP	Validação Ética e Jurídica de Modelos, Programadora Pré-Processamento e Data Augmentation e Anotadora
RENATO ARRUDA	014.910.281-06	TRF3R	Validação Ética e

ROCHA MONTEIRO			Jurídica de Modelos, Curadoria e Anotador
ROBERTO NONATO BARRETO COELHO E SILVA	037.041.364-41	TRF3R	Cientista de Dados e Programador Full-Stack
RODRIGO DE MELO ALMEIDA	263.950.418-51	TRF3R	Anotador
RODRIGO GONÇALVES YUNOGUTHI	318.063.378-60	SJSP	Programador Full-Stack
RODRIGO VIEIRA DA SILVA	411.102.928-05	TRF3R	Engenheiro de IA
ROGÉRIO ANTÔNIO BATISTA DE ARAÚJO	127.550.718-26	SJSP	Programador Full-Stack, Documentação e Apoio Operacional
ROSANA MORAES ZONARO	117.809.008-64	TRF3R	Anotadora
SÉRGIO RICARDO LOZANO	133.626.998-71	SJSP	Cientista de Dados e Programador Pré-Processamento e Data Augmentation
SIMONE MONTEACUTI MARTIN	080.018.428-94	SJSP	Anotadora
SINARA MARIA REIS CHAVES	798.621.956-49	SJSP	Curadoria
VAL EMERSON ARALDI	136.982.178-64	SJSP	Cientista de Dados, Programador Full-Stack e Anotador
VITOR NEVES	081.558.218-89	TRF3R	Anotador

RIBEIRO			

Linha de pesquisa:

a) Soluções para automação dos processos e rotinas de trabalho da atividade judiciária;

b) Soluções de apoio à decisão dos magistrados.

2. INTRODUÇÃO

O potencial inovador da aplicação da Inteligência Artificial à Linguagem Natural no âmbito do Poder Judiciário encontra ambiente fértil na Terceira Região com a digitalização dos processos, implantação do Processo Judicial Eletrônico, criação dos laboratórios de inovação e disponibilidade de tecnologias "estado da arte" em código aberto.

Nesse cenário, pretende-se a utilização de ferramentas de Entendimento de Linguagem Natural (NLU) para a criação de soluções tanto para automação dos processos e rotinas de trabalho da atividade judiciária quanto para o apoio à decisão dos magistrados (itens 1.3.1 e 1.3.2 do Edital 2/2019 do CNJ).

Para tanto, necessária a criação de algorítimos de entendimento de um texto jurídico de uma peça processual, os quais servirão de apoio a tarefas mais específicas, tais como o agrupamento de processos semelhantes, identificação de assuntos para distribuição, auxílio na produção de minutas através de seleção de modelos, pesquisa de Jurisprudência, emissão automática de certidões, sumarização etc.

Esses algorítimos de entendimento de um texto jurídico, considerando a necessidade de auditoria e de facilidade de manutenção quando ocorrerem alterações legislativas e jurisprudenciais, precisam ser modulares.

O primeiro módulo a ser implementado é o "SINARA", o qual, utilizando tecnologias de Named Entity Recognition (NER) seguidas de Relation Extraction (RE), extrairá de um texto jurídico pré-processado dispositivos legais e precedentes citados.

Posteriormente, pretende-se a criação de novos módulos de Extração de Informações (IE), viabilizando soluções mais abrangentes para os problemas mencionados.

Embora estejam sujeitos a alterações, conforme viabilidade, disponibilidade de dados, acurácia e necessidades do projeto, vislumbra-se, como exemplos, módulos para identificar as seguintes informações:

● Partes - muitas matérias na Justiça Federal são específicas de ações com partes determinadas, tais como União, Caixa Econômica Federal, conselhos de classe etc. Indentificá-las pode ser muito útil para o entendimento do processo.
● Pedido e causa de pedir - identificar o pedido e a causa de pedir em petições dos advogados é uma tarefa bem mais complexa, mas que, se alcançada, certamente proporcionará soluções importantes na classificação das demandas.
● Tributos, benefícios previdenciários e outros pontos centrais das matérias na Justiça Federal - além de identificá-los é importante qualificá-los, seja através de “semantic parsing” (solução mais elaborada) ou criando “embeddings” com as palavras próximas (solução imediata).

A primeira solução a ser criada com a utilização desses módulos será a leitura da petição inicial com o objetivo de agrupamento e identificação de assunto.
O projeto encontra um mínimo entregável com a identificação dos dispositivos legais em textos jurídicos e o estudo do potencial de reconhecimento do contexto a partir de seu fundamento legal.

3. PROBLEMA DE PESQUISA

O foco do projeto é a identificação de dispositivos normativos, precedentes e súmulas em textos jurídicos.
Esse módulo poderá ser utilizado para viabilizar ou melhorar a acurácia de outros problemas preditivos, já que muito se pode saber sobre o contexto de uma peça jurídica a partir de seu fundamento legal.
O módulo de identificação dos dispositivos será baseado em "tokens" com implementação não conjunta de algoritmos para reconhecer as entidades (NER) e extrair as relações (RE). O reconhecimento das entidades, por não utilizar embeddings contextualizados, deverá levar em consideração informações de POS e análise sintática extraídos de outra base de dados.

4. JUSTIFICATIVA

A utilização, no âmbito do Poder Judiciário, do potencial inovador da aplicação da Inteligência Artificial à Linguagem Natural encontra um terreno fértil na Justiça Federal da 3.ª Região em razão da digitalização de seu acervo processual, da implantação do Processo

Judicial Eletrônico e da criação dos laboratórios de inovação, bem como em razão da disponibilidade de tecnologias “estado da arte” em código aberto.

Faz-se necessária a busca de soluções para automação dos processos e rotinas de trabalho da atividade judiciária e para o apoio à decisão dos magistrados, conforme preconizado nos itens 1.3.1 e 1.3.2 do Edital 2/2019 do CNJ.

Tais necessidades dependem da criação de algoritmos de entendimento de um texto jurídico constante de uma peça processual, os quais servirão de apoio a tarefas mais específicas, tais como o agrupamento de processos semelhantes, identificação de assuntos para distribuição, auxílio na produção de minutas através de seleção de modelos, pesquisa de Jurisprudência, emissão automática de certidões etc. Deve ser considerada a modularização desses algoritmos de entendimento de um texto jurídico, considerando as necessidades de auditorias e manutenções quando ocorrerem alterações legislativas e jurisprudenciais.

Nesse contexto, propomos um módulo de extração de dispositivos normativos, considerando que muito se pode entender sobre um texto jurídico a partir de seu fundamento legal.

5. OBJETIVOS

O objetivo da pesquisa é a utilização de técnicas tradicionais de reconhecimento de entidades e extração de relações para a identificação dos dispositivos normativos, bem como estudar sua influência na identificação do contexto da peça jurídica. Acreditamos que esse trabalho pode ter impacto significativo em todas as pesquisas futuras sobre o tema.

Havendo tempo, pretendemos substituir os métodos tradicionais de RE e NER por métodos conjuntos (JOINT) baseados em "spans", "transformers" e "embeddings" contextualizados, que hoje alcançam os melhores resultados para a tarefa.

6. METODOLOGIA

O treinamento supervisionado utilizará dados anotados por um time multidisciplinar, capacitado pelo Laboratório de Inteligência Artificial Aplicada da 3ª Região - LIAA-3R, dividido em anotadores e curadores. A base para anotação será gerada por um time de RegEx, responsável pela preparação dos textos extraídos do PJE.

É interessante para o projeto as informações do CODEX do SINAPSES, principalmente para ampliar as formas de treinamento.

A aferição dos resultados será por precisão, recall e F1 para cada entidade do NER bem como para a RE. Deve-se atentar também para que esses índices não estejam viciados devido a um bias do dataset. Por fim, estaremos atentos ao tempo de resposta de cada algoritmo.

7. CRONOGRAMA

ENTREGA	CRONOGRAMA	
	INÍCIO	TÉRMINO
Definição e preparação dos ambientes de pré-processamento de texto e anotações	16/12/2019	19/12/2019
Realização de Workshops sobre fluxos de trabalho, pré-processamento e anotação	02, 08, 09 e 14 e 15/01/2020	16/01/2020
Atividades de pré-processamento	09/01/2020	17/02/2020
Atividades de anotação	21/12/2019	17/02/2020
Atividades de curadoria	17/01/2020	17/02/2020
Implementação do Named-entity recognition (NER)	18/02/2020	23/02/2020
Implementação do extrator de relações	23/01/2020	23/04/2020
Implementação de agrupador	23/04/2020	28/05/2020
Escrever *paper*	28/02/2020	29/05/2020
Submeter projeto no SINAPSES	29/05/2020	29/05/2020

8. REFERÊNCIAS PRELIMINARES

A escolha do tema deriva da monografia abaixo colacionada, que demonstra o avanço na área de NER/RE, superando em muito os métodos tradicionais aplicados antes da explosão do "Deep Learning". A literatura também colabora com possibilidades variadas para treinamento semi-supervisionado, aumentando as possibilidade de se atingir um resultado satisfatório.

Agichtein, E., & Gravano, L. (2000). Snowball: Extracting relations from large plain-text collections. Proceedings of the Fifth ACM International Conference on Digital Libraries.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction from the web. IJCAI ’07: Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India.

Bowen Yu, Zhenyu Zhang, Jianlin Su, Yubin Wang, Tingwen Liu, Bin Wang, Sujian Li. 11-Sep-2019. "Joint Extraction of Entities and Relations Based on a Novel Decomposition Strategy".

Brin, S. (1998). Extracting patterns and relations from the world wide web. WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT ’98.

Bunescu, R. C., & Mooney, R. J. (2005). Subsequence kernels for relation extraction. Neural Information Processing Systems, NIPS 2005, Vancouver, British Columbia, Canada.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao, ‘Relation Classification via Convolutional Deep Neural Network’, in Proc. of COLING 2014, pp. 2335–2344, Dublin, Ireland, (August 2014). Dublin City University and ACL.

Dongxu Zhang and Dong Wang, ‘Relation Classification via Recurrent Neural Network’, CoRR, abs/1508.01006, (2015).

Dixit, K., & Al-Onaizan, Y. (2019). Span-Level Model for Relation Extraction. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy.

Eberts, M., & Ulges, A. (2019) Span-based Joint Entity and Relation Extraction with Transformer Pre-training. arXiv:1909.07755v3.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder, ‘Joint entity recognition and relation extraction as a multi-head selection problem’, Expert Systems with Applications, 114, 34–45, (04 2018).

Markus Eberts, Adrian Ulges. 17-Sep-2019. "Span-based Joint Entity and Relation Extraction with Transformer Pre-training".

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, Luke Zettlemoyer. 20 Mar 2018. "AllenNLP: A Deep Semantic Natural Language Processing Platform"

McDonald, R. (2004). Extracting relations from unstructured text. UPenn CIS Technical Report.

McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y., & White, P. (2005). Simple algorithms for complex relation extraction with applications to biomedical ie. ACL ’05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 491–498). Ann Arbor, Michigan.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, pages 1003–1011. Association for Computational Linguistics.

Stephen Mayhew, Tatiana Tsygankova, Dan Roth. IJCNLP 2019. 31-Aug-2019. "ner and pos when nothing is capitalized".

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd conference on Association for Computational Linguistics (pp. 189–196). NJ, USA.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi, ‘A general framework for information extraction using dynamic span graphs’, in Proc. of NAACL-HLT 2019, volume 1, pp. 3036–3046, Minneapolis, Minnesota, (June 2019). ACL.

Vikas Yadav and Steven Bethard, ‘A Survey on Recent Advances in Named Entity Recognition from Deep Learning models’, in Proc. of the 27th International Conference on Computational Linguistics, pp. 2145–2158, Santa Fe, New Mexico, USA, (August 2018). ACL

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, Yi Chang. 7-Sep-2019. "A Novel Hierarchical Binary Tagging Framework for Joint Extraction of Entities and Relations."