Extração de Metadados para um Grande Arquivo de Decisões Judiciais: uma Abordagem com Inteligência Artificial

João Lima[2], Aerty Santos[4], Eduardo Almeida[4], Juliana Pirovani[2,3,4], and Elias de Oliveira[1,3,4]

[1]Departamento de Arquivologia
[2]Departamento de Informática
[3]Universidade Federal do Espírito Santo
[4]Programa de Pós-Graduação em Informática
[5]Departamento de Computação/Alegre

Resumo

A descrição arquivística costuma ser uma tarefa manual e demorada. Portanto, tende a ser mais estática em seu ciclo de vida da informação. Por outro lado, os usuários mudam com maior frequência a forma com que utilizam os documentos e exigem novos recursos não previstos na descrição inicial dos documentos no arquivo. Esse artigo apresenta uma estratégia para encontrar entidades nomiadas automaticamente, de forma a produzir elementos ricos de metadescrição para um arquivo a ser gerido no ICA-AtoM. Para atingir esse objetivo, aplicamos algumas estratégias de inteligência artificial. Em nossos experimentos, usamos o conjunto de dados de decisões do Tribunal de Justiça de São Paulo em um total de 9.178 e realizamos a extração de alguns novos elementos descritivos. A partir deste corpus, separamos um conjunto menor de documentos relacionados as decisões judiciais dos crimes de Roubo e Furto. Os experimentos mostraram que a abordagem proposta reduziu para menos que 30% o esforço humano para a realização da tarefa com o uso da metodologia proposta baseada em inteligência artificial.

1 Introdução

As ferramentas de recuperação de informações melhoraram na última década. Google, Yahoo, Microsoft e outras empresas oferecem mecanismos de busca poderosos que permitem pesquisar mundialmente por grandes quantidades de documentos quase instantaneamente. Infelizmente, ferramentas semelhantes ainda não estão disponíveis em muitos sistemas de informação usados pelos departamentos de justiça no Brasil. Quando integradas a um sistema de informação institucional local, ferramentas como essas podem aumentar a eficiência e precisão na busca por informações em grandes arquivos de documentos (IZO; OLIVEIRA; BADUE, 2021).

A decisão do juiz é o ato administrativo. Um marco emitido pelo sistema judicial referente a um fato disputado entre partes. Tendo isso em mente, esses documentos são valiosos instrumentos para avaliar o desempenho do sistema judicial. No entanto, para atingir esse objetivo, precisamos construir um melhor sistema de arquivamento para aumentar a acessibilidade desses documentos, vinculando-os a novos pontos de acesso. Os acessos atuais são baseados principalmente no número do processo, no nome das partes e no tipo de assunto em disputa, por exemplo.

A extração de alguns desses metadados pode ser feita diretamente por meio de estratégias baseadas no reconhecimento de padrões. Neste trabalho, vamos além desse tipo de procedimento, extraindo outras informações presentes nesses documentos: por exemplo, o nome das pessoas mencionadas, datas dos fatos mencionados e o tempo aplicado à sentença, apenas para citar algumas dessas informações contidas no texto da decisão. Para peças de informação mais complexas, são necessárias técnicas de Processamento de Linguagem Natural (MANNING; SCHUETZE, 1999). Nossa abordagem segue aquela apresentada por (PIROVANI; OLIVEIRA, 2021). A abordagem daqueles autores usa uma Gramática Local (LG) para cada entidade de interesse de anotação, auxiliando o *Conditional Random Field* (CRF) (SUTTON; MCCALLUM, 2012), outro algoritmo inteligente. O papel de uma LG é inserir o conhecimento do especialista do domínio no processo, já que uma LG é construída de forma eminentemente manual. Algumas abordagens alternativas são vorazes em dados para treinar seus algoritmos (SANTOS; GUIMARAES, 2015), o que pode ser proibitivo em determinado contexto, ou devido ao limitado orçamento da equipe responsável.

São duas as contribuições deste trabalho. Primeiramente, disponibilizamos publicamente um grande conjunto de arquivos de decisões judiciais referentes aos crimes discutidos aqui. Em segundo lugar, organizamos essas decisões de forma que possam ser inseridas em *softwares* populares frequentemente utilizados pelos

arquivistas, como por exemplo o ICA-AtoM[1].

Para apresentar nossa proposta, consideraremos apenas 2 tipos de crimes, cujas decisões foram proferidas em 2022: 1) Furto e 2) Roubo. Para realizar nossos experimentos, consideramos 5.025 documentos.

Este artigo está estruturado da seguinte forma. Na Seção 2, discutimos alguns trabalhos recentes que servirão como base para uma comparação com nossa proposta. A busca por estruturas textuais mais complexas para criar metadados inteligentes permitirá um melhor acesso aos usuários (IZO; OLIVEIRA; BADUE, 2021). Discutimos essa melhoria na Seção 3. Os resultados de nossos experimentos são detalhados na Seção 4, que validam nossa hipótese. Em nossa pesquisa, expandimos a gama de metadados extraídos, resultando em uma notável redução de 30% na carga de trabalho para especialistas humanos em comparação com os processos manuais. Na Seção 5, apresentamos nossas conclusões e pesquisas futuras.

2 Revisão de Literatura

Campos e Oliveira (2015) propõem uma metodologia baseada em gramática local para a identificação automática de nomes de pessoas em textos em português, usando como exemplo o romance Senhora, de José de Alencar. A estratégia adotada, que integra informações contextuais e dicionários, resultou em indicadores de desempenho notáveis, alcançando 99% de recall e 100% de precisão na identificação de 1699 nomes quando aplicada ao livro Senhora. No entanto, ao estender essa abordagem para a classe Tudo a ver (TAV) do conjunto de dados ATribuna, apenas 63 dos 116 nomes existentes foram corretamente identificados, com 9 falsos positivos, resultando em 54% de recall e 88% de precisão. O desempenho inferior nessa classe específica pode ser atribuído à ausência de uma gramática especificamente construída para ela, destacando a necessidade de adaptação da abordagem de acordo com as características únicas de cada conjunto de dados.

Entidades nomeadas identificadas automaticamente em textos têm sido usadas em várias aplicações com o objetivo de reduzir o esforço humano em tarefas manuais e repetitivas. Pirovani, Spalenza e Oliveira (2017), por exemplo, usaram 10 tipos de entidades nomeadas identificadas automaticamente usando CRF+LG em textos educacionais para gerar perguntas, economizando o tempo do professor. Eles encontraram 7195 entidades nomeadas e geraram 6917 perguntas (4175 de preenchimento de lacunas e 2742 de texto livre).

Pirovani, Nogueira e Oliveira (2018) também usaram CRF+LG para encontrar nomes de pessoas em um grande conjunto de dados de um jornal. Os nomes

[1]https://ica-atom.org/

encontrados foram usados para criar uma página web de índice de nomes de pessoas. Essa página permite que o usuário acesse facilmente todos os artigos do jornal onde seus nomes aparecem e vá diretamente ao ponto onde o nome é citado na página do jornal.

Além disso, entidades nomeadas identificadas em textos também são usadas para aumentar a compreensão do assunto do texto e melhorar a qualidade de sistemas com diferentes propósitos. Oliveira et al. (2019) usaram entidades nomeadas como características adicionais para aprimorar a qualidade dos sistemas de avaliação automatizada de redações. Com as características de cada redação respectiva, incluindo as entidades nomeadas, as notas são então inseridas como a variável critério em uma regressão linear. O estudo mostrou a importância de levar em consideração as entidades nomeadas para avaliar a qualidade dos argumentos dos alunos em relação ao tema abordado na redação. Eles obtiveram uma melhoria em duas das cinco dimensões dos critérios de avaliação usados pelo exame ENEM em relação aos resultados de ponta.

Entidades nomeadas químicas foram usadas por Izo et al. (2023) para gerar relatórios inteligentes para a tomada de decisões a partir de textos livres. Eles utilizaram patentes químicas e artigos científicos, extraíram as entidades nomeadas através do método híbrido CRF+LG e, em seguida, aplicaram regras para gerar informações inteligentes. Foram gerados 35 relatórios inteligentes combinando informações de documentos que citam métodos, equipamentos, compostos, elementos e classes químicas similares.

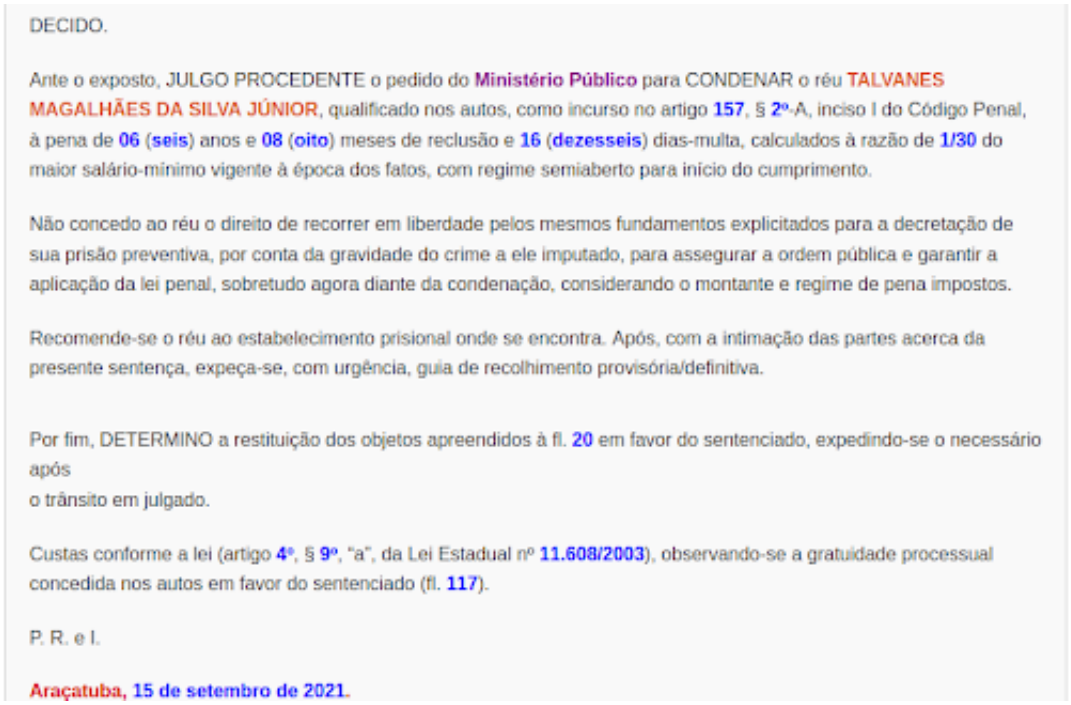
3 A Proposta

Um documento carrega em si muitos elementos textuais que são valiosas informações para alguém, ou alguéns em algum dia. Não podemos prever com certeza, *a priori*, qual palavra ou conjunto de estruturas textuais complexas, entre muitos outros fenômenos linguísticos textuais, serão de uso valioso para um pesquisador, ou um cidadão, ou mesmo para arquivistas. Por essa razão, precisamos de melhores ferramentas e metodologias para reformular rapidamente nossa estrutura de descrição arquivística a fim de atender às necessidades particulares dos usuários do momento.

Novos algoritmos de inteligência artificial estão à nossa disposição para anotar estruturas linguísticas ricas a partir de textos (PIROVANI; OLIVEIRA, 2021). Com base em uma abordagem semelhante ao artigo citado anteriormente, focamos, no presente trabalho, em identificar e anotar alguns elementos de uma coleção de decisões judiciais. Esses elementos comporão os metadados desses documentos ao registrá-los no ICA-AtoM, por exemplo. Com esse objetivo, aprimoramos o algoritmo anterior para identificar o tempo da sentença na decisão judicial,

entre outros elementos discutidos nos resultados, na Seção 4.

Primeiramente, anotamos manualmente 30 decisões para treinar o algoritmo. Durante esse processo manual, marcamos todos os elementos que queremos que o algoritmo aprenda: 1) o número do processo, o documento; 2) a data de emissão da decisão; 3) a comarca; 4) o tribunal; 5) o nome do juiz responsável; 6) o nome do réu, ou réus; e 7) a sentença definitiva. Esta é o tempo da sentença após alguma fração de tempo adicional ou de redução com base nas circunstâncias em que o crime foi cometido. A Figura 1 mostra um exemplo de parte de uma decisão anotada. O réu foi condenado a 6 anos, 8 meses, e uma multa de 16 dias de acordo com 1/30 do maior salário mínimo vigente.



DECIDO.

Ante o exposto, JULGO PROCEDENTE o pedido do **Ministério Público** para CONDENAR o réu **TALVANES MAGALHÃES DA SILVA JÚNIOR**, qualificado nos autos, como incurso no artigo **157**, § **2º**-A, inciso I do Código Penal, à pena de **06** (**seis**) anos e **08** (**oito**) meses de reclusão e **16** (**dezesseis**) dias-multa, calculados à razão de **1/30** do maior salário-mínimo vigente à época dos fatos, com regime semiaberto para início do cumprimento.

Não concedo ao réu o direito de recorrer em liberdade pelos mesmos fundamentos explicitados para a decretação de sua prisão preventiva, por conta da gravidade do crime a ele imputado, para assegurar a ordem pública e garantir a aplicação da lei penal, sobretudo agora diante da condenação, considerando o montante e regime de pena impostos.

Recomende-se o réu ao estabelecimento prisional onde se encontra. Após, com a intimação das partes acerca da presente sentença, expeça-se, com urgência, guia de recolhimento provisória/definitiva.

Por fim, DETERMINO a restituição dos objetos apreendidos à fl. **20** em favor do sentenciado, expedindo-se o necessário após
o trânsito em julgado.

Custas conforme a lei (artigo **4º**, § **9º**, "a", da Lei Estadual nº **11.608/2003**), observando-se a gratuidade processual concedida nos autos em favor do sentenciado (fl. **117**).

P. R. e I.

Araçatuba, 15 de setembro de 2021.

Figura 1: Um exemplo de decisão judicial anotada.

Após ter um conjunto de documentos anotados manualmente, os algoritmos propostos utilizaram esses documentos anotados para aprender a anotar automaticamente os documentos que ainda não foram anotados por especialistas.

4 Experimentos e Resultados

O problema de anotação de texto ainda é um desafio para as comunidades científicas. Um dos problemas é o custo para treinar um algoritmo a aprender a anotar

uma grande quantidade de documentos de arquivo. Estes algoritmos geralmente precisam de um grande conjunto de dados previamente rotulados (LI et al., 2022). Isso significa, portanto, um grande esforço de anotações humanas e validações cruzadas.

Propomos uma abordagem em que o custo para treinar nosso algoritmo é mais barato do que a maioria das alternativas da literatura. Em outras palavras, nosso algoritmo pode aprender rapidamente, e seu desempenho de qualidade é estatisticamente comparável aos algoritmos literários de referência. Para capturar alguns dados de treinamento, desenvolvemos um aplicativo de software, baseado no Moodle[2], para permitir que especialistas humanos insiram suas anotações em dados que serão usados para treinamento. O sistema solicita a um grupo de usuários selecionados que anotem uma amostra de documentos usando uma linguagem baseada em XML para marcar entidades nomeadas que podem ser encontradas nos documentos (por exemplo, Figura 1). Isso é feito até que amostras suficientes sejam acumuladas para treinar o algoritmo. Os resultados de acertos em anotações apresentados por outros autores são muito promissores (COLOMBO; OLIVEIRA, 2022; PIROVANI; OLIVEIRA, 2021).

5 Conclusão

Poropõe-se neste trabalho uma abordagem com o uso de inteligência artificial para acelerar a tarefa de descrição arquivística. Mais do que apenas prestar atenção aos elementos textuais tradicionais, como por exemplo o número do processo, os nomes das partes e o tipo do assunto em disputa, nossa abordagem busca no documento inteiro por entidades nomeadas para melhorar a capacidade de acesso dos usuários aos seus documentos.

Aumentar a qualidade dos metadados arquivísticos com novas e mais complexas expressões textuais, como entidades nomeadas, permite melhorias na busca em sistemas de informação, como já discutido pelo autores Izo, Oliveira e Badue (2021). Portanto, o usuário é amplamente beneficiado pela incorporação dessa tecnologia em um sistema de informação. No trabalho atual, estudamos os marcos iniciais para fornecer essas funcionalidades aos documentos de decisões judiciais. Com a incorporação dessa funcionalidade pode-se inclusive inferir-se a classificação dos documentos baseando-se apenas nos valores das sentenças.

Um trabalho futuro que planejamos realizar é a extração de palavras, ou expressões, que expliquem o aumento ou diminuições das sentenças pelos juízes. Ou seja, busca-se assim as razões e explicações dos motivos que sustentaram a atribuição das sentenças. Por outro lado, mecanismos como os que buscamos jogará

[2]http://moodle.org/

luzes a que novas metodologias possam ser estudadas para estabelecer diretrizes e critérios claros para os juízes seguirem ao determinar suas sentenças.

Referências

CAMPOS, J.; OLIVEIRA, E. Extração de Nomes de Pessoas em Textos em Português: uma Abordagem Usando Gramáticas Locais. In: *Computer on the Beach 2015*. Florianópolis, SC: SBC, 2015. *Best paper award. Disponível em: <https://hal.archives-ouvertes.fr/hal-01134971>*.

COLOMBO, C.; OLIVEIRA, E. Intelligent Information System for Extracting Knowledge from Pharmaceutical Package Inserts. In: *XVIII Simpósio Brasileiro de Sistemas de Informação (SBSI)*. Florianópolis, SC: SBC, 2022.

IZO, F.; OLIVEIRA, E.; BADUE, C. Named Entities as a Metadata Resource for Indexing and Searching Information. In: *21$^{\underline{th}}$ International Conference on Intelligent Systems Design and Applications – (ISDA)*. On the WWW: Springer International Publishing, 2021. v. 418, p. 838–848.

IZO, F. et al. An Intelligent Report Generator for Chemical Documents. In: *XIX Brazilian Symposium on Information Systems*. Curitiba, Brazil: Association for Computing Machinery, 2023. p. 276–283.

LI, Q. et al. A Survey on Text Classification: From Traditional to Deep Learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM New York, NY, v. 13, n. 2, p. 1–41, 2022.

MANNING, C. D.; SCHUETZE, H. *Foundations of Statistical Natural Language Processing*. 1st. ed. New York, NY: The MIT Press, 1999.

OLIVEIRA, E. et al. The Influence of NER on the Essay Grading. In: SPRINGER. *19$^{\underline{th}}$ International Conference on Intelligent Systems Design and Applications: Intelligent Systems Design and Applications*. Delhi, India: Springer International Publishing, 2019. p. 102–113.

PIROVANI, J.; NOGUEIRA, M.; OLIVEIRA, E. Indexing Names of Persons in a Newspaper Large Dataset. In: *13$^{\underline{th}}$ International Conference on the Computational Processing of Portuguese (PROPOR)*. Canela, RS: Springer, 2018. v. 11122.

PIROVANI, J.; OLIVEIRA, E. Studying the Adaptation of Portuguese NER for Different Textual Genres. *The Journal of Supercomputing*, Springer International Publishing, p. 1–17, 2021.

PIROVANI, J.; SPALENZA, M.; OLIVEIRA, E. Geração Automática de Questões a Partir do Reconhecimento de Entidades Nomeadas em Textos

Didáticos. In: *XXVIII Simpósio Brasileiro de Informática na Educação (SBIE)*. Ceará, CE: SBC, 2017. p. 1147–1156.

SANTOS, C. N.; GUIMARAES, V. Boosting Named Entity Recognition with Neural Character Embeddings. In: *Proceedings of the Fifth Named Entities Workshop, ACL 2015*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015. p. 25–33.

SUTTON, C.; MCCALLUM, A. Conditional Random Fields: An Introduction. *Foundations and Trends® in Machine Learning*, Elsevier, v. 4, p. 267–373, 2012. ISSN 1935-8237.