A comparative analysis of text embedding approach to extract named entities in Portuguese legal documents

Hyan H. N. Batista[1], André C. A. Nascimento[1], Rafael Ferreira Melo[1], Péricles B. C. Miranda[1], Isabel W. S. Maldonado[2], José L. M. Coelho Filho[2]

[1] Departamento de computação
Universidade Federal Rural do Pernambuco (UFRPE)
Recife – PE – Brasil
[2] NESS Law, São Paulo, Brasil

{hyan.batista,andre.camara,rafael.mello,pericles.miranda}@ufrpe.br,
{iwanderley,leo}@ness.com.br

Abstract. *The initial petition is one of the most important components of a civil litigation process. Automating the analysis of these documents might reduce the time necessary for the postulatory phase's conclusion. The parties qualification body is the section in which are exposed the informations about the entities involved in the process. This paper suggests the employment of named entity extraction techniques on the problem of information extraction and recognition on initial petitions. With this in mind, was created a part description body corpora extracted from Brazilian courts. Seven BiLSTM-CRF models with distinct combinations of vector representations of words were trained, evaluated, and compared to investigate their effects on the performance of an algorithm with that architecture and, in this way, improve the recognition of legal entities in legal texts. Unlike other works based on BiLSTM-CRF for NER tasks in the legal domain, this research emphasizes not the architectures employed, but rather the text representation methods used. The experiments performed with the developed corpus show that the stacking of character, word, and pooled FLAIR embeddings is the preferred combination to extract the best possible performance from BiLSTM-CRF hybrid models.*

Resumo. *A petição inicial é um dos componentes mais importantes de um processo civil, de modo que a automatização da análise desses documentos pode diminuir o tempo necessário para que se cumpra a fase postulatória. O corpo de qualificação das partes, por sua vez, é a seção nesse documento onde são expostas as informações a respeito das entidades envolvidas no processo. Este artigo propõe o uso de técnicas de extração de entidades nomeadas no problema de identificação e extração de informações em petições iniciais. Para tal, foi produzida uma base de dados composta por corpos de qualificação das partes de petições iniciais extraídas de processos advindos de tribunais brasileiros. Foram treinados, avaliados e comparados sete modelos BiLSTM-CRF com combinações distintas de representações vetoriais de palavras, a fim de se investigar seus efeitos na performance de um algoritmo com essa arquitetura e, dessa forma, aprimorar o reconhecimento de entidades jurídicas em textos legais. Ao contrário de outros trabalhos baseados em BiLSTM-CRF para tarefas de NER no domínio jurídico, esta pesquisa dá ênfase não às arquiteturas empregadas, mas sim aos métodos de representação de texto usados. Os*

experimentos executados com o corpus desenvolvido mostram que o empilhamento de incorporações de caracteres, palavras e pooled FLAIR embeddings é a combinação preferível para extrair-se o melhor desempenho possível de modelos híbridos BiLSTM-CRF.

1. Introdução

O setor judiciário se baseia na troca constante de documentos entre diferentes partes de um processo. Neste contexto, a petição inicial é uma peça jurídica que inicia os processos, por isso ela é um documento essencial em qualquer processo civil. Esta peça jurídica contém várias informações cruciais para o andamento do caso, como dados relacionados a nomes dos autores e réus do processo, número de documentos de identificação (e.g., CPF, CNPJ), endereços eletrônicos ou físicos, entre outros. Entretanto, por tratar-se de um texto essencialmente narrativo, essas informações não são estruturadas e sua escrita é carregada de jargões jurídicos, o que dificulta o uso de técnicas computacionais para as tarefas de coleta, armazenamento e distribuição de informações.

Nesse contexto, as técnicas de processamento de linguagens natural (PLN) oferecem, justamente, um meio de transformar dados textuais, como o conteúdo de uma petição inicial, em dados estruturados, permitindo que esses dados possam ser coletados, armazenados e distribuídos [Deng and Liu 2018]. Mais especificamente, para a extração de informações de petições iniciais, uma das tarefas mais importantes é o reconhecimento de entidades nomeadas (do inglês *Named Entity Recognition* (NER) [Li et al. 2020a]) a partir de textos puros, visto que é através dela que a maior parte dos dados, como nome de pessoas, organizações e CNPJs, são extraídos.

Técnicas PLN e de aprendizagem de máquina (AM) têm se mostrado excelentes ferramentas para o reconhecimento de entidades nomeadas [Ritter et al. 2011] [Li et al. 2020b], inclusive no domínio jurídico [Wang et al. 2020]. Em [Luz de Araujo et al. 2018], os autores exploram a aplicação de arquiteturas híbridas na tarefa de classificação de sequências de palavras utilizando uma base de dados própria, o LeNER-br. Vários trabalhos relatam que as técnicas que vem alcançando melhores resultados para aplicações de NER em diferentes contextos são as baseadas em redes profundas [Li et al. 2020a]. Isto também é relatado em trabalhos específicos do contexto jurídico [Leitner et al. 2020]. Contudo, para utilizar os algoritmos de redes profundas é necessário utilizar camadas de *embedding* para representar o texto. Por outro lado, é importante destacar que embora este tema tenha recebido bastante atenção nos últimos anos, o número de estudos realizados para textos no idioma português ainda é bastante limitado. Isto se deve em parte à dificuldade de acesso a bases de dados anotadas. Além disso, a maior parte dos trabalhos nesta temática dá ênfase às arquiteturas e algoritmos de NER, não analisando de maneira isolada os métodos de representação de texto usados.

Diante deste contexto, este trabalho propõe uma análise sistemática dos efeitos da escolha de diferentes representações vetoriais de palavras em modelos de NER baseados em aprendizagem de máquina. Mais especificamente, os experimentos foram realizados usando diferentes arquiteturas de embeddings, mas tomando por base um modelo do estado da arte em NER, a rede profunda híbrida BiLSTM-CRF (*Bidirectional Long-Short Term Memory-Conditional Random Fields*) [Lample et al. 2016]. Diversas combinações de representações distintas foram analisadas, a fim de identificar a técnica, ou combinação

de técnicas, que maximiza a taxa de acerto de algoritmos de NER no contexto de textos jurídicos brasileiros. As principais contribuições deste artigo são: 1) Treino, avaliação de comparação da performance de modelos BiLSTM-CRF configurados com diferentes técnicas de representação de palavras como uma solução para o problema de reconhecimento de entidades nomeadas jurídicas em corpos de qualificação das partes. 2) Anotação de uma base de dados de acesso restrito para reconhecimento de entidades nomeadas em língua portuguesa em um contexto jurídico fechado: classificação e etiquetagem de sequências de palavras em um corpo de qualificação das partes.

2. Trabalhos Relacionados

[Luz de Araujo et al. 2018] apresenta um corpus para reconhecimento de entidades nomeadas em documentos legais brasileiros intitulado LeNER-br. Esse *dataset* contém seis tipos de entidades: pessoa, jurisprudência, tempo, localização, legislação e organização. Para estabelecer uma base comparativa, os autores realizaram os experimentos em outro conjunto de dados em língua portuguesa, o Paramopama [Menezes et al. 2019]. O LeNER conta com 70 documentos, 10.392 sentenças e aproximadamente 318.100 *tokens*.

Os autores utilizaram uma arquitetura híbrida LSTM-CNN (*Long-Short Term Memory-Convolutional Neural Network*) assim como descrita em [Mendonça Jr et al. 2016], denominada ParamopamaWNN. O outro modelo trata-se de uma BiLSTM-CRF (*Bidirectional Long-Short Term Memory-Conditional Random Fields*) com incorporações à nível de caractere, uma arquitetura primeiramente introduzida como solução para o problema de etiquetagem de sequências em [Lample et al. 2016]. Os resultados obtidos demonstraram a superioridade da arquitetura híbrida BiLSTM-CRF como sendo a mais eficiente na tarefa de NER, não só no *dataset* Paramopama, mas também no *dataset* proposto pelos autores. As conclusões obtidas estão de acordo com trabalhos anteriores que também demonstram a eficiência da arquitetura híbrida para tarefas de NER [Yadav and Bethard 2019]. É importante ressaltar que os tipos de entidades anotados no LeNER-Br, em especial a entidade PESSOA, não correspondem ao padrão de entidades que precisam ser extraídas de corpo de qualificação das partes de petição inicial, visto que entidades como CPF, CNPJ, RG e OAB não estão presentes na base.

[Sousa and Del Fabro 2019] introduz uma base de textos jurídicos obtidos da consulta pública ao Supremo Tribunal Federal brasileiro, o ITD (*Iudicium Textum Dataset*). A base conta com mais de 40 mil acordões, 48 mil votos e 39 mil relatórios identificados de acordo com o seu ministro redator. Embora essa base possa ser usada para a execução de diversas tarefas de PLN ela não se encaixaria no contexto do problema de extração das partes de uma petição inicial, pois, como sustentado por [Giorgi and Bader 2020], manter o desempenho de um modelo, que foi treinado em uma certa base, estável ao aplicá-lo em uma outra de propósito semelhante é um dos principais desafios do NER. Petições iniciais possuem uma estrutura bastante diferente quando contrastada com acordões e relatórios.

[Leitner et al. 2020], por sua vez, descreve um corpus construído para tarefas de reconhecimento de entidades nomeadas em decisões de tribunais federais alemães. A base de dados produzida nesse trabalho é composto por cerca de 67.000 sentenças, as quais possuem aproximadamente dois milhões de *tokens*. O corpus contém dezenove classes semânticas, são elas: pessoa, juíz, advogado, país, rua, paisagem, organização, compa-

nhia, instituição, corte, marca, lei, ordenança, norma legal europeia, regulação, contrato, decisão de tribunal e literatura legal. Essas classes estão distribuição ao longo de 54.000 entidades manualmente anotadas sobre a base de dados produzida. Embora o *dataset* construído pelos autores possua entidades que coincidem com as entidades que desejaria-se reconhecer em um corpo de qualificação das partes, o fato deles estar na língua alemã torna-o inutilizável no treinamento de serviços de NER para tribunais brasileiros.

Os experimentos realizados por [Leitner et al. 2020] consideraram os modelos que até então compunham o estado-da-arte usando duas variações de anotações, refinadas (*fine-grained*) e não refinadas (*coarse-grained*). Os algoritmos escolhidos para testar o corpus foram CRF-F (com atributos), CRF-FG (com atributos e *gazetteers*), CRF-FGL (com atributos, *gazetteers* e *lookup*) e BiLSTM-CRFs com incorporações de palavras pré-treinadas [Reimers et al. 2014]: BiLSTM-CRF [Huang et al. 2015], BiLSTM-CRF+ com incorporação de caracteres de BiLSTM [Lample et al. 2016] e BiLSTM-CNN-CRF com incorporação de caracteres de CNN [Ma and Hovy 2016]. Indo em consonância com o trabalho de [Luz de Araujo et al. 2018], os melhores resultados, em ambos os tipos de anotação, refinada e não refinada, foram obtidos pela arquitetura BiLSTM-CRF, mais especificamente, considerando *embeddings* de palavras e caracteres.

Enquanto trabalhos anteriores concentraram-se na extração de entidades jurídicas de escopo geral, esta pesquisa dá ênfase em classes específicas que são encontradas dentro do corpo de qualificação das partes em uma petição inicial. Além disso, conseguiu-se estudar mais amplamente o efeito que tipos diferentes de incorporações possuem nos resultados de modelos baseados em aprendizagem de máquina e híbridos no que se refere à precisão, cobertura e F_1 score.

3. Metodologia

3.1. Base de dados

A produção da base de dados considerou um conjunto inicial de 10.000 petições iniciais em formato digital, associadas a metadados que incluíam os nomes dos autores, réus e advogados. Como as petições estavam no formato de pdf, o conteúdo textual de cada documento foi extraído utilizando um serviço de reconhecimento ótico de caracteres (OCR). Os textos extraídos foram segmentados em seções para que se pudesse selecionar apenas a seção das partes, que se encontra os dados pessoais relacionados aos envolvidos no processo. Esse processo resultou na extração 7.966 textos, os quais foram posteriormente segmentados em parágrafos.

A anotação das entidades foi feita de maneira semi-automática, a partir dos metadados associados a cada processo, bem como de conjuntos de regras explícitas, na forma de regras de expressões regulares. As regras produzidas envolviam aspectos regulares da qualificação das partes, como estado civil (e.g., ”solteiro”, ”solteira”, ”casado”, ”casada”, etc.), nacionalidade (”brasileiro”, ”brasileira”, ”bras.”, etc.) e regras associadas a documentos de identificação (e.g., RG, CPF e CNPJ). Ao final, foram produzidos 23.630 sentenças, as quais foram submetidas a uma análise de qualidade, considerando apenas sentenças que possuíam pelo menos um nome de autor e um de réu. Esse processo de filtragem resultou em um conjunto de dados composto por 640 sentenças. Cada sentença contém informações relacionadas à: nome de pessoas, nacionalidade, estado civil, número de RG, número de CPF e CNPJ, e número da OAB dos advogados envolvidos.

Figura 1. Distribuição das classes no conjunto de treino.

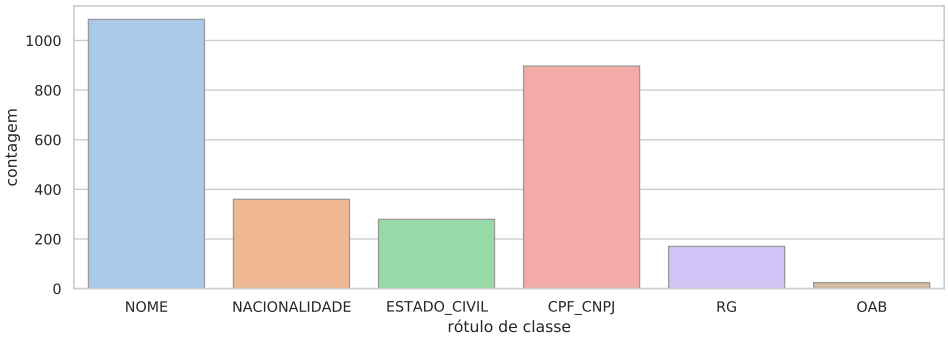
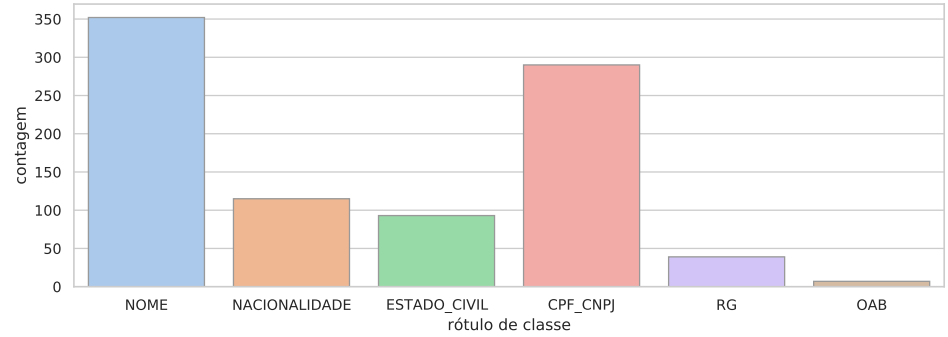


Figura 2. Distribuição das classes no conjunto de teste.



As sentenças resultantes foram então separadas aleatoriamente em conjuntos de treino e teste para avaliação do modelo, sendo o primeiro deles composto por 480 sentenças, e o segundo, 160.

A Figura 1 e a Figura 2 mostram a distribuição das classes nos conjuntos de treino e teste, respectivamente. Pode-se observar que as classes encontram-se desbalanceadas, sendo NOME a classe mais frequente com 1437 amostras, enquanto a menos frequente, OAB com apenas 31. Isso se deve à natureza heterogênea dos exemplos. A frequência de entidades do tipo RG, NACIONALIDADE e ESTADO CIVIL, por exemplo, é diretamente proporcional à presença de nomes de pessoas físicas. Além disso, como as anotações foram feitas através do uso de expressões regulares, podem acontecer falhas e algumas entidades acabam não sendo marcadas. Os textos contém ao todo 132.689 *tokens* e, após a separação dos dados, o conjunto de treino ficou com 100.379 *tokens* e o de teste, com 32.310.

3.2. Arquiteturas de embedding avaliados

Para este trabalho, foram avaliadas as seguintes arquiteturas de *embedding*: *word embedding*, *character embedding*, *FLAIR embedding* [Akbik et al. 2018] e *Pooled Contextualized Embedding* [Akbik et al. 2019].

Existem vários modelos pré-treinados de *word embeddings* para uso [Rudkowsky et al. 2018]. Contudo, para a arquitetura de *word embedding* avaliada aqui selecionou-se o *Glove embeddings* [Pennington et al. 2014], devido ao seu amplo

uso na literatura [Khatri et al. 2020] [Wang et al. 2018, Gaikwad and Haribhakta 2020]. A Glove é que um algoritmo não supervisionado que tem como propósito obter representações vetoriais para palavras. Para isso, o Glove mapeia palavras em um espaço vetorial onde a distância entre as palavras é diretamente proporcional a sua similaridade semântica [Pennington et al. 2014]. As palavras do conjunto de treino que não estão no dicionário do Glove são denominadas OOV (em inglês, out-of-vocabulary). O algoritmo lida com essas palavras as posicionando aleatoriamente no espaço vetorial [Pennington et al. 2014]. A função da próxima arquitetura de *embedding* usada neste trabalho, *character embedding* é utilizar um rede neural convolucional e uma única dimensão para dar a essas palavras OOV representações numéricas adequadas [Du et al. 2016].

FLAIR embeddings ou *contextual string embeddings* é tipo de *word embedding* relativamente recente na literatura [Akbik et al. 2018]. Esse tipo de *embedding* são treinadas sem nenhuma noção explicita das palavras, portanto, fundamentalmente modela palavras como sequências de caracteres e, além disso, são contextualizadas pelo texto ao redor dessas sequências, o que significa que a mesma palavra terá diferentes representações vetoriais dependendo de seu uso em contexto [Akbik et al. 2018]. Abordagens baseadas em caracteres como essa, contudo, enfrentam dificuldades para produzir *embeddings* significativas se um cadeia de caracteres rara é usada em um contexto subespecificado [Akbik et al. 2019]. Para lidar com esse incoveniente, [Akbik et al. 2019] propõe um método, denominado *Pooled contextualized embedding*, no qual agrega-se dinamicamente *embeddings* contextualizadas de cada cadeia única de caracteres que é encontrada. O método, então, usa uma operação de *pooling* para destilar uma representação global para a palavra a partir de todas as instâncias contextualizadas.

Nesta pesquisa, para avaliar o efeito desses *embeddings*, testa-se as seguintes combinações: *word embedding* isolado (WORD), *character embedding* isolado, *word embedding* e *character embedding* (CHAR + WORD), *word e flair embedding* (WORD + FLAIR), *character, word* e *flair embedding* (CHAR + WORD + FLAIR), *word* e *pooled contextualized embedding* (WORD + POOL) e, por fim, *character, word* e *pooled contextualized embedding* (CHAR + WORD + POOL).

3.3. Bidirectional Long-Short Term Memory-Conditional Random Fields

A arquitetura *Bidirectional Long-Short Term Memory-Conditional Random Fields* (BiLSTM-CRF) vem sendo amplamente usada em vários trabalhos de reconhecimento de entidades nomeadas como uma linha de base para análise de desempenho de modelos e conjuntos de dados [Luz de Araujo et al. 2018] [Leitner et al. 2020]. Ela nasceu na união de um método baseado em redes neurais, a Memória de Curto e Longo Prazo (LSTM), e um outro que se baseia em grafos probabilísticos discriminativos, o Campo Aleatório Condicional de Cadeia Linear (CRF).

Uma LSTM (em inglês, *Long-Short Term Memory*) é uma variante específica de redes neurais recorrentes que busca mitigar o problema do desaparecimento de gradientes explosivos na modelagem de dependência de curto e longo prazo de uma sequência de caracteres [Hong and Lee 2020]. Se uma arquitetura LSTM é bidirecional, então ela processa a cadeia de caracteres de entrada em ambas as direções, ou seja, para frente (em inglês, *forward*) e para trás (em inglês, *backward*), usando duas LSTMs. Um CRF (em

inglês, *Linear-chain Conditional Random Field*), por sua vez, é uma classe de modelos de grafos probabilísticos discriminativos que descreve a probabilidade conjunta de *P*(**y**—**x**) de todos os rótulos estruturados **y** com relação à estrutura de um grafo não-orientado, dado um conjunto de entradas **x** [Hong and Lee 2020]. Segundo [Hong and Lee 2020], CRF é uma técnica amplamente usada em vários problemas de rotulação de sequências como também em BioNER (Reconhecimento de Entidade Nomeadas Biomédicas) por meio da aplicação da propriedade de Markov de primeira ordem na rotulação de sequência de saída. A junção desses métodos resulta em uma BiLSTM-CRF (em inglês, *Bidirectional Long-Short Term Memory-Conditional Random Fields*). Essa arquitetura é tipicamente composta por quatro camadas: a camada de *token embedding*, camada *token embedding* BiLSTM, camada de enlace (em inglês, binding layer) e, por fim, a camada de CRF [Hong and Lee 2020].

3.4. Métodos de avaliação

Para avaliar modelos de NER é necessário se ter um *dataset* anotado no qual passa-se executá-lo e, em seguida, comparar seus resultados com seu padrão-ouro. Essa comparação é feita, nesta pesquisa, ao nível das frases e as pontuações são contadas apenas para as sequências detectadas que combinam tanto em posicionamento na frase, quanto em tipo de classe.

Tendo isso em mente, de acordo com [Mohit 2014], as métricas avaliação comumente usadas são a precisão e cobertura. A cobertura é a porcentagem de entidades nomeadas anotadas que o modelo é capaz de detectar. Por outro lado, a precisão mede a porcentagem de entidades nomeadas anotadas que coincidem com as anotações do *dataset* padrão-ouro. Essas relações podem ser expressas matematicamente pelas expressões:

$$cobertura = \frac{TP}{TP + FN} \qquad precisão = \frac{TP}{TP + FP} \tag{1}$$

Nessas expressões, TP refere-se a verdadeiro positivos, FN refere-se a falso negativos e FP refere-se a falso positivos. Uma terceira métrica de avaliação bastante utilizada é a *F_1 score* apresentada na Equação 2. A métrica F_1 tem sido a avaliação *de facto* para o reconhecimento de entidades nomeadas, devido à sua simplicidade e generalidade [Mohit 2014].

$$F_1 \text{ score} = \frac{2 \times precisão \times cobertura}{precisão + cobertura} \tag{2}$$

4. Resultados

Nesta seção são apresentados os resultados acerca dos experimentos realizados para esta pesquisa. A performance dos modelos treinados foi avaliada sob a perspectiva de diferentes métricas (precisão, cobertura e *F_1 score*). Com relação à recursos de hardware e software, os experimentos foram executados no *Gradient*, uma plataforma de *cloud computing* focada em inteligência artificial e aprendizagem de máquina. A instância utilizada foi a *Free-GPU*, a qual fornecia acesso a uma GPU (*Grafical Processing Unit*), 30GB de memória RAM (Random Access Memory) e 8 CPUs (Central Processing Unity). A versão do Python empregada foi a 3.8 e do *FLAIR*, a 0.8.0.

Tabela 1. *F_1 score* das arquiteturas testadas para cada classe anotada no *dataset*.

Representação Textual	Nome	Nacion.	Estado Civil	RG	OAB	CPF/CNPJ
CHAR	0.5894	0.9739	0.9721	0.5352	0.2222	0.9347
WORD	0.8187	0.9791	0.8449	0.7368	0.4615	0.9023
CHAR + WORD	0.8322	0.9873	0.9945	0.8687	0.4286	0.9526
WORD + FLAIR	0.8427	**0.9957**	0.9891	**0.9278**	**0.4615**	0.9832
CHAR + WORD + FLAIR	0.8535	**0.9957**	0.9891	0.9184	**0.4615**	0.9831
WORD + POOL	**0.8649**	**0.9957**	**0.9945**	0.9200	0.4286	0.9832
CHAR + WORD + POOL	0.8563	**0.9957**	**0.9945**	0.9109	**0.4615**	**0.9853**

A Tabela 1 apresenta o F_1 *score* para cada uma das classes do *dataset*. Pode-se perceber que de modo geral o empilhamento de *embeddings* produz resultados superiores aos apresentados pelas arquiteturas de *embedding* isoladas. Para mais, a Tabela 3 mostra a média aritmética dos F_1 *scores* para cada um dos rótulos de classe. Seus dados evidenciam ainda mais a conclusão anterior mostrando um desempenho quase que gradual das pontuações conforme empilha-se os *embeddings*.

Na Tabela 2 são apresentados, por sua vez, os resultados para todos os métodos testados ao longo dos experimentos. Todos os métodos possuem fundamentalmente a mesma arquitetura: uma BiLSTM-CRF. Entretanto, procurou-se nos experimentos executados nesta pesquisa ressaltar não a performance de determinadas técnicas, sejam essas baseadas em regras, probabilidade, aprendizagem de máquina ou híbridas, mas sim a influência que determinadas representações possuem no desempenho de um modelo BiLSTM-CRF. O uso de tal arquitetura foi uma consequência do estudo da literatura sobre reconhecimento de entidades nomeadas em domínios específicos que demonstram claramente a superioridade desta arquitetura em relação à outras quando se trata de tarefas de etiquetagem e classificação de sequências de palavras. Como foi estabelecida uma mesma arquitetura base para todos os testes, a Tabela 2 apresenta, portanto, como a performance dessa arquitetura varia conforme o tipo de representação textual usado. Para esta pesquisa foram selecionadas as seguintes incorporações, ou combinação de incorporações: caractere, palavra [Huang et al. 2015], palavra e caractere [Lample et al. 2016], palavra e *FLAIR embeddings* [Akbik et al. 2018]; caractere, palavra e *FLAIR embeddings* e, por fim, palavra e *pooled FLAIR embeddings* [Akbik et al. 2019]. Ao observar-se os dados apresentados na Tabela 2, é notório que a combinação de incorporações de caracteres, palavras e *pooled FLAIR embeddings*, salvo duas exceções, para as entidades RG e OAB, é a que produz os melhores resultados com relação à precisão, cobertura e F_1 *score* para as entidades jurídicas anotadas na base de dados. Para a classe Nacionalidade, as combinações CHAR + WORD, WORD + POOL e CHAR + WORD + POOL adquiriram scores iguais. Situação semelhante se repete para a classe Estado Civil onde as combinações CHAR + WORD + FLAIR, WORD + POOL e CHAR + WORD + POOL obtiveram, resultados perfeitos nas três métricas analisadas. No que se refere às classes RG e OAB, a combinação com maior F_1 *score* foi a CHAR + WORD. As combinações CHAR + WORD + FLAIR e CHAR + WORD + POOL, no entanto, ficaram apenas um pouco atrás. Com relação à classe OAB, contudo, a distância entre os resultados dessas combinações e a CHAR + WORD foram bem mais significativas, as quais, inclusive, foram as maiores para essa classe.

Tabela 2. Precisão, cobertura e *F_1 score* dos modelos BiLSTM-CRF treinados e suas incorporações.

Classe	*Embedding(s)*	Precisão	cobertura	*F_1 score*
Nome	CHAR	0.7362	0.4915	0.5894
Nome	WORD	0.8164	0.8210	0.8187
Nome	CHAR + WORD	0.8130	0.8523	0.8322
Nome	WORD + FLAIR	0.8333	0.8523	0.8427
Nome	CHAR + WORD + FLAIR	0.8547	0.8523	0.8535
Nome	WORD + POOL	0.8661	0.8636	0.8649
Nome	CHAR + WORD + POOL	0.8411	0.8722	0.8563
Nacionalidade	CHAR	0.9912	0.9573	0.9739
Nacionalidade	WORD	0.9590	1.0000	0.9791
Nacionalidade	CHAR + WORD	0.9750	1.0000	0.9873
Nacionalidade	WORD + FLAIR	0.9915	1.0000	0.9957
Nacionalidade	CHAR + WORD + FLAIR	0.9915	1.0000	0.9957
Nacionalidade	WORD + POOL	0.9915	1.0000	0.9957
Nacionalidade	CHAR + WORD + POOL	0.9915	1.0000	0.9957
Estado Civil	CHAR	0.9886	0.9560	0.9721
Estado Civil	WORD	0.8229	0.8681	0.8449
Estado Civil	CHAR + WORD	0.9891	1.0000	0.9945
Estado Civil	WORD + FLAIR	0.9785	1.0000	0.9891
Estado Civil	CHAR + WORD + FLAIR	0.9785	1.0000	0.9891
Estado Civil	WORD + POOL	0.9891	1.0000	0.9945
Estado Civil	CHAR + WORD + POOL	0.9891	1.0000	0.9945
RG	CHAR	0.7917	0.4043	0.5352
RG	WORD	0.7292	0.7447	0.7368
RG	CHAR + WORD	0.8269	0.9149	0.8687
RG	WORD + FLAIR	0.9000	0.9574	0.9278
RG	CHAR + WORD + FLAIR	0.8824	0.9574	0.9184
RG	WORD + POOL	0.8679	0.9787	0.9200
RG	CHAR + WORD + POOL	0.8519	0.9787	0.9109
OAB	CHAR	0.5000	0.1429	0.2222
OAB	WORD	0.5000	0.4286	0.4615
OAB	CHAR + WORD	0.4286	0.4286	0.4286
OAB	WORD + FLAIR	0.5000	0.4286	0.4615
OAB	CHAR + WORD + FLAIR	0.5000	0.4286	0.4615
OAB	WORD + POOL	0.4286	0.4286	0.4286
OAB	CHAR + WORD + POOL	0.5000	0.4286	0.4615
CPF/CNPJ	CHAR	0.9328	0.9367	0.9347
CPF/CNPJ	WORD	0.8893	0.9156	0.9023
CPF/CNPJ	CHAR + WORD	0.9315	0.9747	0.9526
CPF/CNPJ	WORD + FLAIR	0.9791	0.9873	0.9832
CPF/CNPJ	CHAR + WORD + FLAIR	0.9831	0.9831	0.9831
CPF/CNPJ	WORD + POOL	0.9791	0.9873	0.9832
CPF/CNPJ	CHAR + WORD + POOL	0.9792	0.9916	0.9853

Tabela 3. Média aritmética dos *F_1 scores* de cada classe para cada uma das etiquetas de classe.

Incorporação	Média
CHAR	0,7046
WORD	0,7905
CHAR + WORD	0,8439
WORD + FLAIR	0,8667
CHAR + WORD + FLAIR	0,8668
WORD + POOL	0,8645
CHAR + WORD + POOL	**0,8674**

Tendo em vista os resultados exibidos na Tabela 3, é notório que o empilhamento de incorporações em diferentes níveis, no geral, traz pontuações maiores nas métricas comumente aplicadas na avaliação de sistemas de reconhecimento de entidades nomeadas (precisão, cobertura e F_1 *score*). Além disso, é perceptível que o emprego das incorporações introduzidas em [Akbik et al. 2018] e [Akbik et al. 2019] na composição de vetores de palavras, em quase todas as situações, aprimora os resultados gerados. Pode-se dizer que, diante dos fatos apresentados, que em tarefas de classificação de sequências de palavras em um âmbito legal, com a existência de vocabulário especializado, o uso de de incorporação de caracteres, palavras e *pooled FLAIR embeddings* é a melhor empilhamento possível, quando comparada com os seus pares, para se aplicar em arquiteturas BiLSTM-CRF.

5. Conclusão

A pesquisa apresentada neste artigo tem como objetivo estudar o uso de diferentes técnicas de incorporação em modelos BiLSTM-CRF para o problema de reconhecimento de entidades nomeadas no corpo de qualificação das partes de petições iniciais. Através de experimentos realizados em documentos jurídicos reais de acesso restrito, os resultados demonstram que o empilhamento de incorporações de caracteres, palavras e *pooled FLAIR embeddings* é a configuração preferível para se extrair a melhor performance possível de modelos híbridos BiLSTM-CRF. Dessa forma, esta pesquisa contribui com um entendimento mais profundo sobre os efeitos de tais técnicas no desempenho de algoritmos de *machine learning* sob a perspectiva da precisão, cobertura e F_1 *score* dos modelos produzidos. Os artefatos confeccionados neste trabalho podem ser utilizados por cientistas de dados que atuam no setor jurídico para expandir o ecossistema de inteligência artificial de suas empresas e, assim, ampliar a diversidade de produtos fornecidos. Para trabalhos futuros, sugere-se a expansão da base de dados, tanto no que concerne seu número de amostras quanto com relação às etiquetas de classe empregadas.

Referências

Akbik, A., Bergmann, T., and Vollgraf, R. (2019). Pooled contextualized embeddings for named entity recognition. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 724–728.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Deng, L. and Liu, Y. (2018). *Deep learning in natural language processing*. Springer.

Du, L., Li, X., Liu, C., Liu, R., Fan, X., Yang, J., Lin, D., and Wei, M. (2016). Chinese word segmentation based on conditional random fields with character clustering. In *2016 International Conference on Asian Language Processing (IALP)*, pages 258–261. IEEE.

Gaikwad, V. and Haribhakta, Y. (2020). Adaptive glove and fasttext model for hindi word embeddings. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 175–179.

Giorgi, J. M. and Bader, G. D. (2020). Towards reliable named entity recognition in the biomedical domain. *Bioinformatics*, 36(1):280–286.

Hong, S. and Lee, J.-G. (2020). Dtranner: biomedical named entity recognition with deep learning-based label-label transition model. *BMC bioinformatics*, 21(1):1–11.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Khatri, A. et al. (2020). Sarcasm detection in tweets with bert and glove embeddings. *arXiv preprint arXiv:2006.11512*.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, pages 260–270.

Leitner, E., Rehm, G., and Moreno-Schneider, J. (2020). A dataset of German legal documents for named entity recognition. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, (Section 5):4478–4485.

Li, J., Sun, A., Han, J., and Li, C. (2020a). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

Li, J., Sun, A., Han, J., and Li, C. (2020b). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

Luz de Araujo, P. H., de Campos, T. E., de Oliveira, R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11122 LNAI:313–323.

Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Mendonça Jr, C., Barbosa, L. A., Macedo, H. T., and São Cristóvão, S. (2016). Uma arquitetura híbrida lstm-cnn para reconhecimento de entidades nomeadas em textos naturais em língua portuguesa. *XIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). SBC*.

Menezes, D. S., Milidiú, R. L., and Savarese, P. (2019). Building a massive corpus for named entity recognition using free open data sources. *Proceedings - 2019 Brazilian Conference on Intelligent Systems, BRACIS 2019*, pages 6–11.

Mohit, B. (2014). Named entity recognition. In *Natural language processing of semitic languages*, pages 221–245. Springer.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Reimers, N., Eckle-Kohler, J., Schnober, C., Kim, J., and Gurevych, I. (2014). Germeval-2014: Nested named entity recognition with neural networks.

Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1524–1534.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., and Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, 12(2-3):140–157.

Sousa, A. W. and Del Fabro, M. D. (2019). Iudicium textum dataset uma base de textos jurıdicos para nlp. In *Brazilian Symposium on Databases*, pages 1–11.

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., and Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20.

Wang, Z., Wu, Y., Lei, P., and Peng, C. (2020). Named entity recognition method of brazilian legal text based on pre-training model. In *Journal of Physics: Conference Series*, volume 1550, page 032149. IOP Publishing.

Yadav, V. and Bethard, S. (2019). A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. *arXiv*.