Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Reconhecimento de Entidades Nomeadas para Conteúdo Publicado em Diários Oficiais com Base em uma Abordagem de Supervisão Fraca

Lucélia Vieira Mota

Dissertação apresentada como requisito parcial
para conclusão do Mestrado em Informática

Orientador
Prof. Dr. Thiago de Paulo Faleiros

Brasília
2023

Vieira Mota, Lucelia
VM917r Reconhecimento de Entidades Nomeadas para Conteúdo Publicado em Diários Oficiais com Base em uma Abordagem de Supervisão Fraca / Lucelia Vieira Mota; orientador Thiago de Paulo Faleiros. -- Brasília, 2023.
82 p.

Dissertação(Mestrado em Informática) -- Universidade de Brasília, 2023.

1. aprendizado fracamente supervisionado. 2. reconhecimento de entidade nomeada. 3. supervisão fraca. 4. Diário oficial do Distrito Federal. 5. dados anotados. I. de Paulo Faleiros, Thiago , orient. II. Título.

Dedicatória

Dedico esta dissertação a minha filha Maria Luíza e a minha família por todo o amor e incentivo.

Dedico ao meu Superintendente do SERPRO e amigo, Giordanni Paiva por todo apoio e incentivo.

Dedico também ao meu orientador, Prof. Dr. Thiago de Paulo Faleiros, pelos ensinamentos, orientação, paciência e confiança.

Agradecimentos

Agradeço

A Deus, pela vida, persistência, perseverança, fé e por acreditar que seria possível.

Aos meus pais, Celso e Marilucia por sempre me apoiar e incentivar a levar os estudos com muito afinco, além de serem uma fonte de amor e carinho inesgotável.

A minha filha, Maria Luíza, pelos abraços carinhosos nos momentos de desespero, pelas águas e cafés trazidos quando era difícil me ausentar por alguns minutos de frente do computador.

Ao meu companheiro, Nilson, pela parceria e apoio nos dias ruins, e por me propiciar as condições necessárias para chegar até o final.

Aos meus irmãos, Luciene e Washington Bruno, por sempre fazerem de tudo para me manter motivada diante dos momentos difíceis, me confortando e me encorajando a seguir em frente.

Ao amigo Luis Furasté, que nunca cogitou a palavra impossível, e que me fez acreditar que deveria trilhar esse caminho e me abriu as portas necessárias para chegar até aqui.

Ao Prof. Dr. Thiago de Paulo Faleiros pela orientação e direcionamento ao longo dessa jornada, além dos ensinamentos e experiências, a confiança em mim depositada ao me receber após um processo difícil e doloroso de substituição de orientador. É imensurável a quantidade de conhecimento adquirido por mim nesses mais de 2 anos juntos.

Aos meus amigos pelos momentos de alegria, por todo apoio e pela compreensão da ausência parcial nos últimos anos.

Aos meus colegas do projeto KnEDLe, em especial ao Micael Felipe e Matheus Stauffer, que contribuíram com ensinamentos relevantes para o desenvolvimento desse trabalho.

Aos meus chefes, Carlos Rodrigo e Marcelo Pita, pela compreensão pela minhã não disponibilidade em todos os projetos e pelo apoio para completar todas as etapas.

Aos meus colegas de SERPRO, em especial, ao Alísio, Daniel e Cristiane, pelas palavras positivas, pela compreensão das minhas falhas nos dias de sono e cansaço. E por não me deixarem para trás, e sempre com muita paciência e carinho me puxavam junto, muitas vezes cobrindo lacunas importantes.

Resumo

O Reconhecimento de Entidade Nomeada em português é uma tarefa desafiadora, especialmente em textos formais e oficiais, como Licitações e Contratação Pública. A anotação manual desses textos é cara, demorada e requer conhecimento específico no domínio. Este estudo propõe a criação de um *corpus* anotado de Licitação e Contratação Pública utilizando métodos de supervisão fraca (SF). Estes métodos empregam técnicas de aprendizado de máquina semi-supervisionados para extrair entidades nomeadas de textos não anotados. A aplicação dos métodos fracamente supervisionados, combinando o uso de anotações fracas e funções de rótulo de conhecimentos heurísticos, correspondência de palavras e modelos de aprendizado de máquina pré-treinados desempenham um papel crucial na tarefa de NER, especialmente em cenários nos quais grandes quantidades de dados anotados não estão disponíveis, são caros de obter ou são impraticáveis de rotular manualmente. Assim, adotou-se uma metodologia que possibilitou a geração de um *corpus* de Licitação e Contratação Pública e a validação desse *corpus* com um *corpus* formal anotado manualmente. Para validação deste estudo foram realizados experimentos com modelos CRF, Bi-LSTM-CNN e SF para NER. Os resultados do modelo Bi-LSTM, treinado com os dados provenientes da supervisão fraca, demonstraram um desempenho significativo, atingindo um *F1 Score* médio de 84,3%, contra apenas 0,756% da base ouro. Notavelmente, o destaque foi para o treinamento do Bi-LSTM-CNN com os dados gerados pela supervisão fraca do ato extrato de contrato, alcançando um impressionante *F1 Score* de 96%, superando os 95% obtidos com os dados da base ouro. No entanto, o cenário mais desafiador foi observado no contexto do extrato de convênio, onde a aplicação das FR de supervisão fraca resultou em um *F1 Score* de apenas 47%, em comparação com os 66,9% alcançados pelo CRF sobre a base ouro, acredita-se que esses resultados foram afetados devido a pouca quantidade exemplos no *corpus*. Os resultados obtidos demonstram que a combinação de NER e SF produz um *corpus* de alta qualidade com menos esforço que a anotação manual. Assim, é possível afirmar que o mecanismo de programação do de dados da SF é uma ferramenta promissora para a geração de corpora anotados em português, especialmente em domínios específicos como Licitação e Contratação Pública. Ela acelera o desenvolvimento de ferramentas de NER, reduzindo o tempo e o custo da anotação manual. Este estudo pode ser aplicado para melhoria da ferramenta de NER para o português, desenvolvimento de sistemas de informação para o setor público e extração de informação de documentos de Licitação e Contratação Pública.

Palavras-chave: aprendizado fracamente supervisionado, reconhecimento de entidade nomeada, supervisão fraca, dados anotados, diários oficiais do distrito federal

Abstract

Named Entity Recognition in Portuguese is a challenging task, especially in formal and official texts, such as Bidding and Public Procurement. Manual annotation of these texts is expensive, time-consuming, and requires specific domain knowledge. This study proposes the creation of an annotated corpus of Bidding and Public Procurement using weak supervision methods (WS). These methods employ semi-supervised machine learning techniques to extract named entities from unlabeled texts. The application of weakly supervised methods, combining the use of weak annotations and label functions of heuristic knowledge, word matching, and machine learning, plays a crucial role in the NER task, especially in scenarios where large amounts of annotated data are not available, are expensive to obtain, or are impractical to label manually. Thus, a methodology was adopted that enabled the generation of a corpus of Bidding and Public Procurement and the validation of this corpus with a manually annotated gold standard corpus. To validate this study, experiments were conducted with CRF, Bi-LSTM-CNN e WS para NER. The results of the Bi-LSTM-CNN model, trained with weak supervision data, showed significant performance, achieving an average F1 Score of 84.3%, compared to only 0.756% of the gold standard base. Notably, the highlight was the training of the Bi-LSTM-CNN with weak supervision data for the contract extract act, achieving an impressive F1 Score of 96%, surpassing the 95% obtained with the gold standard data. However, the most challenging scenario was observed in the context of the covenant extract, where the application of weak supervision functions resulted in an F1 Score of only 47%, compared to the 66.9% achieved by CRF on the gold standard base, it is believed that these results were affected due to the small number of examples in the corpus. The results obtained demonstrate that the combination of NER and WS produces a high-quality corpus with less effort than manual annotation. Thus, it can be stated that WS data programming is a promising tool for generating annotated corpora in Portuguese, especially in specific domains such as Bidding and Public Procurement. It accelerates the development of NER tools, reducing the time and cost of manual annotation. This study can be applied to improve NER tools for Portuguese, develop information systems for the public sector, and extract information from Bidding and Public Procurement documents.

Keywords: weakly supervised learning, named entity recognition, weak supervision, annotated data, official gazettes of the Federal District

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

Bi-LSTM *Bi-direcional Long Short Term Memory*. v, vi, ix–xi, 2, 10, 12–14, 24, 26, 27, 34, 40, 44–46, 49–51, 66–72

BoW *Bag of words*. 10

CNN *Convolutional Neural Network*. v, vi, 13, 34, 40, 44–46

Convi1D 1D convolution layer. 45, 46

CPO Corpus Padrão Ouro. 2, 4, 29, 41–47, 49, 51–53

CRF *Linear-chain Conditional Randon Fields*. v, vi, x, xi, 2, 10–12, 23, 24, 26, 27, 34, 40, 43, 44, 46, 49–51, 66–72

DL *Deep Learning*. 10, 40

DODF Diário Oficial do Distrito Federal. 2–4, 29, 30

DP desvio padrão. 66–72

FR *Funções de Rótulo*. v, 1–5, 17–20, 22, 23, 25, 26, 28, 31–34, 38–41, 44, 46, 49, 51–54

GDF Governo do Distrito Federal. 32–34, 37, 38, 49–51

HMM *Hidden Markov Model*. 11, 20, 23, 24

IBGE Instituto Brasileiro de Geografia e Estatistica. 19

KnEDLe *Knowledge Extraction from Documents of Legal content*. iv, 4, 5, 29, 41

L-BFGS Limited Memory Broyden–Fletcher–Goldfarb–Shanno. 43

LSTM *Long-Short Term Memory*. ix, 12, 13, 23, 26, 45, 46

NER *Named Entity Recognition*. v, vi, 1–6, 9–12, 16, 20, 22, 24, 28, 30–32, 39–47, 49, 52–54, 66

PDF Portable Document File. 29

PLN Processamento de Linguagem Natural. 2, 6, 8, 9

RNN *Recurrent Neural Network*. ix, 12, 13

SERPRO Serviço Federal de Processamento de Dados. iii, iv

SF Supervisão Fraca. v, ix–xi, 2–5, 16, 17, 22, 23, 28, 30, 31, 34, 39–44, 46, 47, 49–54, 66–72

Skweak Weak Supervision NLP. 17, 28, 31, 34, 38

TALLOR Tagging with Learnable Logical Rules. 26

TCDF Tribunal de Contas do Distrito Federal. 2, 31

TXT Text Document File. 29

Capítulo 1

Introdução

Com a explosão do interesse de vários segmentos do mercado na adoção de técnicas de aprendizado de máquina, aumentou significativamente a necessidade de dados de treinamento anotados. Para se alcançar bons resultados de classificação é necessário ter conjuntos maciços de dados de treinamento cuidadosamente anotados (Zhou, 2017). Entretanto, a tarefa de anotar uma quantidade suficiente de amostras em um prazo factível para serem utilizados em aplicações de tempo real representa um caminho crítico para o sucesso do projeto (Zhou, 2017). Uma vez que a construção dessas bases de dados demandam muito esforço manual, devido ao grande número de exemplos a serem anotados e da necessidade de um conhecimento especializado para atribuição correta dos rótulos.

Nesse sentido, a anotação de dados manual pode apresentar um custo elevado, além de possíveis impedimentos no processo de treinamento, tais como o tempo necessário para realizar a tarefa manual de criação do *corpora* e os desafios intrínsecos a essa tarefa de NER (Bach et al., 2019). Para criar um cenário ainda mais desafiador, pode-se pensar no contexto no qual uma organização precise implementar vários modelos que devem ser constantemente iterados e treinados para se adaptar as constantes mudanças reais. Nesse caso, a anotação manual se torna insustentável até mesmo para uma grande organização (Tok et al., 2022).

A aplicação de métodos fracamente supervisionados com rótulos incompletos, inexatos e imprecisos (Tok et al., 2022) surge como uma alternativa a esses desafios, ao oferecer uma abordagem flexível e adaptável. Por exemplo, em conjuntos de dados pequenos, onde a qualidade da rotulação é alta, estratégias como o aprendizado ativo, aprendizado semi-supervisionado e transferência de aprendizado podem ser empregadas, com os dois últimos dispensando intervenção humana na curadoria do modelo (Tok et al., 2022). Por outro lado, em conjuntos de dados com alta densidade de anotações e volume considerável, pode-se adotar métodos supervisionados inexatos, permitindo que o algoritmo faça previsões com base no contexto da entidade. Em cenários onde a anotação é escassa ou possui um alto nível de abstração, a abordagem de aprendizado supervisionado fraco (impreciso) surge como uma alternativa, combinando o uso de anotações fracas e funções de rótulo FR. Aqui, expressões regulares e heurísticas desempenham um papel crucial na tarefa de NER (Tok et al., 2022). Essa diversidade de métodos proporciona uma ampla gama de estratégias para lidar com diferentes contextos e requisitos de anotação, garantindo a eficácia e a adaptabilidade do processo de rotulação em NER.

Assim, a adoção das técnicas das tarefas de Processamento de Linguagem Natural

(PLN), como Reconhecimento de Entidades Nomeadas do inglês *Named Entity Recognition* (NER) combinada com a abordagem de aprendizado supervisionado fraco (impreciso), especialmente a SF por meio da programação dos dados, surge como uma oportunidade viável para anotar as entidades automaticamente, e assim reduzir a necessidade de anotação manual (Zhou, 2017). A tarefa de NER foca no processo de encontrar, extrair e classificar entidades nomeadas em linguagem de texto natural (Luz de Araujo et al., 2018). Para Tok et al. (2022), o aprendizado supervisionado fraco é mecanismo que reúne uma coleção de técnicas de aprendizado de máquina semi-supervisionado que treina os modelos usando anotações menos precisas, em vez das anotações extensas e precisas normalmente usadas no aprendizado supervisionado tradicional. Ou seja, por meio do mecanismo da programação de dados da SF, é possível criar várias FR de uma forma organizada, as quais são utilizadas para anotar cada exemplo identificado, e dessa forma, criar uma base de dados que contenha informações "mais fracas" do que dados totalmente supervisionados (Tok et al., 2022) e dessa forma esses dados podem ser usados para treinar um modelo (Tok et al., 2022).

1.1 Problema e motivação

Segundo Tok et al. (2022), a existência de dados anotados de alta qualidade tornou-se um diferencial para o mercado e a comunidade acadêmica, pois são fundamentais para a construção de aplicativos que exigem uma etapa de aprendizado de máquina com base em uma tarefa de NER. A composição desses dados forma o que se conhece como Corpus Padrão Ouro (CPO), que segundo Wißler et al. (2014) é definido como "um *corpus* anotado manualmente e revisado por especialistas". No entanto, nem sempre é possível ter especialistas disponíveis para anotação manual, além do esforço e tempo necessários para realizar essa tarefa.

Modelos de NER frequentemente rodam sobre arquiteturas convolucionais, redes neurais recorrentes ou Bi-LSTM na maioria das vezes complementada por uma camada CRF (Chiu and Nichols, 2015) (Lison et al., 2020). Essas arquiteturas requerem a existência de um grande corpora anotado, como, por exemplo, Ontonotes (Weischedel, 2013) e ConLL 2003 (Sang and De Meulder, 2003). Abordagens como aprendizado ativo ou transferência de aprendizado surgem como alternativas para minimizar o impacto da falta de um *corpus* com muitos exemplos anotados. No entanto, em um contexto onde não há dados anotados, a aplicação dessas abordagens se torna desafiadora (Lison et al., 2020).

No Brasil, os auditores do Tribunal de Contas do Distrito Federal (TCDF) enfrentam um cenário árduo para extrair os textos do Diário Oficial do Distrito Federal (DODF). Este periódico serve como uma fonte primordial de informações sobre os atos oficiais praticados pelo governo, fornecendo um panorama abrangente das atividades administrativas. Com uma vasta gama de detalhes e uma frequência regular de publicação, o DODF abrange uma ampla variedade de temas, incluindo, mas não se limitando a, licitações, contratos e outras deliberações governamentais. O propósito principal dessa publicação é promover a transparência nas ações do governo, permitindo que funcionários públicos e profissionais interessados acessem informações cruciais, como a oficialização de determinado ato ou detalhes sobre uma ação específica, como a data e o órgão envolvido.

No entanto, a natureza abrangente e linguagem natural dos Diários Oficiais tornam o processo de extração e anotação de informações uma tarefa complexa. Com uma grande

diversidade de assuntos presentes em um único documento, a identificação e anotação de informações estruturadas demandam um esforço considerável. O conteúdo desses diários é predominantemente voltado para o domínio da administração pública, o que amplia ainda mais o desafio, exigindo um trabalho minucioso e diário para extrair e anotar informações relevantes dos diversos departamentos governamentais.

A tabela 1.1 ilustra exemplos de entidades contidas nos DODF:

Tabela 1.1: Exemplos de Entidades de Licitação e Contratos

Entidade	Tipo Entidade	Ato de Licitação Pública
116/2018	numero licitacao	Aviso de Suspensão
pregão eletronico	modalidade licitação	Extrato de Contrato
00410-00024534/2017-65	processo gdf	Extrato de Contrato
R$ 153.374,62	valor estimado da contratação	Extrato de aditamento
20/11/2018	data de assinatura	Extrato de Convênio

Nesse contexto desafiador, aplicação dos métodos fracamente supervisionados, como as FR da abordagem de Supervisão Fraca (SF), emergem como uma solução promissora. Esse mecanismo emprega técnicas de aprendizado de máquina para extrair automaticamente rótulos a partir de texto não rotulado, por meio do desenvolvimento de FR do tipo correspondência por palavra, conhecimento heurístico (regras) e modelos treinados por aprendizado de máquina (Lison et al., 2020). Essa proposta simplifica significativamente o processo de anotação, permitindo que especialistas concentrem seus esforços na curadoria das anotações geradas, ao invés de gastar tempo valioso na extração manual.

Assim, este estudo propõe a criação de um *corpus* anotado de Licitação e Contratação Pública empregando técnicas de aprendizado fracamente supervisionados para extrair entidades nomeadas de textos não rotulados. Essa abordagem não apenas agiliza o processo de anotação, como também abre novas possibilidades para a automação de tarefas essenciais no contexto da administração pública.

1.2 Questão de pesquisa e hipótese

Este trabalho assume como questão de pesquisa:

- Como gerar modelos automatizados e precisos para a tarefa de NER, particularmente quando o contexto em questão possui poucos dados anotados disponíveis em português e os conjuntos de dados existentes exibem características contextuais específicas que requerem o envolvimento de pessoal especializado para anotação de dados?

Como resposta a esta pergunta, será investigada a hipótese de que as técnicas da abordagem de SF podem contribuir para a rápida geração de dados anotados e para criação de modelos com qualidade semelhante aos modelos criados com dados anotados

por humanos. Esse estudo de caso é o objeto de pesquisa de um projeto maior pertencente ao projeto KnEDLe[1], o qual fornecerá as bases de dados e os recursos para a avaliação dos objetivos definidos.

1.3 Objetivos

Esta pesquisa propõe a criação de um *corpus* anotado de Licitação e Contratação Pública automatizadamente, utilizando métodos de aprendizado fracamente supervisionados, seguida pela comparação do desempenho de modelos de aprendizagem de máquina treinados sobre a base CPO de Licitação e Contratação Pública com modelos treinados sobre a base criada por SF. E, com isso, explorar o potencial de métodos de SF para aprimorar a tarefa de NER no *corpus* doDODF. A fim de nortear o alcance do objetivo desta pesquisa foram definidos os seguintes objetivos específicos:

1. Investigar métodos de geração de entidades nomeadas a partir das técnicas de aprendizado fracamente supervisionados.
2. Selecionar as entidades de licitação e contrato a serem utilizadas nos experimentos.
3. Descrever os padrões ou características das entidades de licitação e contrato selecionadas.
4. Implementar as FR do tipo conhecimento heurístico, correspondência por palavra e aprendizado de máquina.
5. Criar base de dados anotados utilizando os métodos da abordagem de SF.
6. Treinar modelos de NER sobre o CPO.
7. Treinar modelos de NER sobre a base gerada pela abordagem de SF.

1.4 Contribuições

As principais contribuições desta dissertação são as seguintes:

1. Apoio na condução e execução das atividades do projeto KnEDLe.
2. Validação do CPO dos atos de licitação e contrato.
3. Geração da base de dados de SF dos atos de licitação e contrato.
4. Aplicação prática dos conceitos de SF derivados deste estudo para a construção de uma API NER para ser comercializada pelo Serviço Federal de Processamento de Dados do Governo Federal.

Alguns resultados de publicação científica foram alcançados durante a realização deste trabalho, conforme segue:

[1]http://nido.unb.br/

- Mota, L. V, Faleiros, T., Lima, A. Pereira, A., Borges, V., Queiroz, A. (2023). *Named Entity Recognition For Content Published In Government Gazettes Based On A Weak Supervision Approach*, 18 Iberian Conference on Information Systems and Technologies, CISTI 2023, Portugal, June 19, 2023, Proceedings.
- Ferreira, I., Lopes, L., Faleiros, T., Garcia, L., Borges, V., Queiroz, A., Mota, L. V (2023). *A Tool For Retrieving The Life-Cycle of Public Procurement Processes*. 18 Iberian Conference on Information Systems and Technologies, CISTI 2023, Portugal, June 19, 2023, Proceedings.

Além de um artigo da construção da base ouro dos atos de licitação e contratos que está sendo elaborado em parceria com a equipe do projeto KnEDLe para envio a uma revista internacional, conforme segue:

- Lima, A. Pereira, A., Borges, V., Queiroz, A. (2023), Mota, L. V, Faleiros, T.. *A Gold Standard Corpus for Named Entity Recognition of Bidding Processes and Contract Excerpts Publications of a Government Gazette.*

1.5 Organização da Dissertação

Este capítulo inaugura a compreensão do problema de anotação de dados no contexto brasileiro e na língua portuguesa, delineando a motivação e os objetivos desta pesquisa. No Capítulo 2, são apresentados os fundamentos teóricos essenciais para uma compreensão mais profunda dos experimentos conduzidos e delineados nas seções subsequentes. Em seguida, o Capítulo 3 oferece uma revisão da literatura relevante, destacando os estudos e abordagens relacionados à Supervisão Fraca (SF) aplicada a tarefa de Reconhecimento de Entidades Nomeadas (NER). No Capítulo 4, é detalhada a metodologia adotada para o desenvolvimento da pesquisa, incluindo a implementação das FR necessárias. Já o Capítulo 5 descreve as diretrizes empregadas na execução e validação dos experimentos, oferecendo uma visão abrangente do processo experimental. No Capítulo 6, são apresentados os resultados obtidos, acompanhados por uma breve análise e discussão. Por fim, o Capítulo 7 encerra esta dissertação com uma análise aprofundada dos resultados alcançados, além de indicar possíveis direções para trabalhos futuros, consolidando assim o estudo realizado.

Capítulo 2

Fundamentação Teórica

Neste capítulo, são delineados os conceitos fundamentais relacionados ao escopo desta pesquisa. Inicialmente, será apresentada uma visão geral do processo de Licitação e Contratos, contextualizando sua relevância para o estudo em questão. Em seguida, será explorado os fundamentos da área de Processamento de Linguagem Natural (PLN). Posteriormente, são discutidos conceitos essenciais relativos à tarefa de Reconhecimento de Entidade Nomeada (NER), incluindo modelos utilizados para essa finalidade, princípios de aprendizado de máquina e uma introdução à abordagem das técnicas de aprendizado fracamente supervisionados.

2.1 Processo de Contratação Pública

O processo de contratação pública é o processo público para a realização de compras e de obras públicas, as quais são juridicamente norteadas pela Lei de Licitações e Contratos Administrativos, conhecida como a nova lei de licitações, sob o número 14.133/2021 [1] em substituição às leis 8.666/93 [2] .

O processo de contratação pública, segundo abstração da Lei N. 14.333, pode ser divido em três grandes etapas:

[1]Disponível em `http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/L14133.htm`
[2]Disponível em `http://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm/`

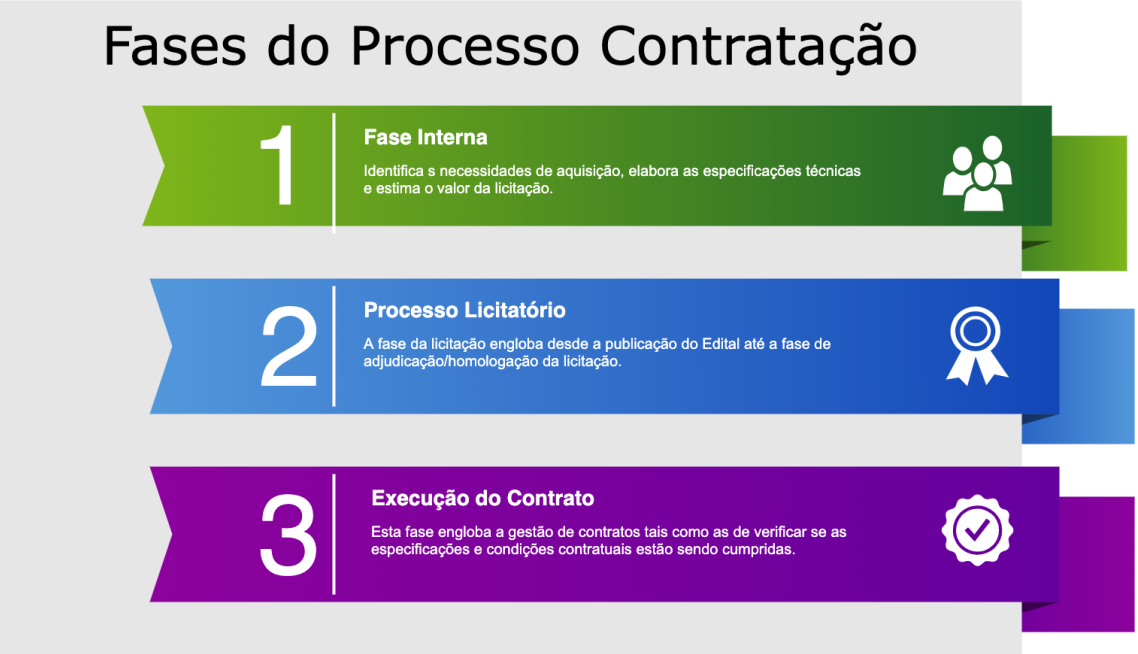


Figura 2.1: Fases do Processo de Contratação conforme interpretação da Lei N. 14.333/2021

Conforme o capítulo I, Art. 12, Inciso VI da Lei N. 14.333/2021, o processo de contratação pública deverá observar:

> "os atos serão preferencialmente digitais, para permitir que sejam produzidos, comunicados, armazenados e validados por meio eletrônico;"

Em cumprimento ao disposto acima, o Tribunal de Contas da União do DF (o órgão do DF) publica os atos do processo de contratação no Diário Oficial da União. Os principais atos publicados são:

- Aviso de Suspensão de Licitação
- Aviso de Revogação/Anulação de Licitação
- Extrato de Contrato ou Convênio
- Extrato de aditamento contratual
- Aviso de Abertura de Licitação

A Figura 2.2 mostra uma publicação no Diário Oficial do DF de um Extrato de Contrato.

CONTRATO PARA AQUISIÇÃO DE BENS PELO DISTRITO FEDERAL nº 09/2022 - PMDF/DSAP, NOS TERMOS DO PADRÃO nº 07/2002. PROCESSO 00054-00145536/2021-05.

O Distrito Federal, por meio do Departamento de Saúde e Assistência ao Pessoal, representado por CORONEL QOPM JORGE MARCOS XAVIER DA SILVA, na qualidade de Chefe do Departamento de Saúde e Assistência ao Pessoal da PMDF, com delegação de competência prevista nas Normas de Execução Orçamentária, Finanças, e Contábil do Distrito Federal, daqui em diante denominado CONTRATANTE e a Empresa J L SILVA - COMERCIO LTDA - CNPJ 40.273.957/0001-26, com sede em RUA DO MERCADO, 06 CENTRO, BURITICUPU/MA, CEP: 65.393-000, Telefone: (98) 98436-5777, e-mail: jozafox3@gmail.com, representada por JEOZADAQUE LIRA SILVA, RG nº ***4518 - SSP/GO, CPF nº 028.***.***-03, na qualidade de CONTRATADA, objetivando a aquisição de Tiras reagentes para determinação de glicemia: Accu-Chek (Performa) e Gtech lite, consoante especifica do Termo de Referência (Doc. SEI nº 75763255) e da Proposta (Doc. SEI nº 77199332). A entrega dos objetos processar-se-á de forma integral em 30 (trinta) dias, contados a partir da retirada/recebimento da respectiva Nota de Empenho, conforme especificação contida no Termo de Referência (Doc. SEI nº 75763255) e na Proposta (Doc. SEI nº 77199332), facultada sua prorrogação nas hipóteses previstas no § 1º, art. 57 da Lei nº 8.666/93, devidamente justificada por escrito e previamente autorizada pela autoridade competente para celebrar o Contrato. A despesa correrá à conta da seguinte Dotação Orçamentária: I – Unidade Orçamentária: 73901; II – Programa de Trabalho: 28845090300FM0053; III – Natureza da Despesa: 3.3.90.30.36; IV – Fonte de Recursos: 151. O empenho tem o valor de R$ 3.265,00 (três mil duzentos e sessenta e cinco reais), conforme Nota de Empenho nº 2022NE000299 (Doc. SEI nº 82361406), emitida em 17 de março de 2022, na modalidade ordinário. O contrato terá vigência de 90 (noventa) dias, a contar da data de sua assinatura. JORGE MARCOS XAVIER DA SILVA, Chefe.

Figura 2.2: Exemplo de publicação no Diário Oficial do DF de um Extrato de Contrato

A Tabela dos Atos e Entidades praticados pelo Tribunal de Contas do DF, no contexto da contratação pública, pode ser visualizada no A.

Os atos praticados são específicos ao contexto da administração pública brasileira, no idioma português e são publicados diariamente no diário oficial do DF [3] . Dentre os problemas encontrados, está a dificuldade em se obter dados rotulados de forma célere e com boa qualidade.

2.2 Processamento de Linguagem Natural

Uma das áreas da inteligência artificial é o PLN que consiste no desenvolvimento de modelos computacionais para a realização de tarefas que dependem de informações expressas em alguma linguagem natural. Muitas tarefas podem ser desenvolvidas dentro deste campo, dentre as quais se encontram as ferramentas usadas no nosso dia-a-dia como, tradutores automatizados, chatbots, filtros de spam ou corretores gramaticais. De acordo com Russell and Norvig (2010) as principais áreas de PLN atualmente são:

- Análise de Sentimentos: usada para interpretar e classificar se a reação a um texto, frase ou documento foi positiva, negativa ou neutra.
- Modelos de Linguagem: predição da próxima palavra ou letra de um texto. Modelos atuais conseguem gerar textos complexos a partir de poucas palavras de modo que o resultado é quase imperceptivelmente artificial.
- Classificação de Textos: associar um texto a uma categoria pré-definida, por exemplo, é possível associar o texto de cada e-mail à categoria Spam ou não Spam.
- Perguntas e Respostas: normalmente com base em um texto de referência, um modelo é treinado para responder perguntas.
- Reconhecimento da fala: a partir de um áudio interpretado é possível traduzir a fala em texto.
- Outros: Reconhecimento de entidades nomeada, sumarização, extração de relacionamento, *etc.*

[3] Disponível em `https://www.dodf.df.gov.br/`

2.3 Reconhecimento de Entidade Nomeada

Algumas das principais tarefas realizadas dentro do PLN estão ligadas a semântica e a sintaxe do texto, como, por exemplo, o NER. O uso de estruturas semânticas e sintáticas pela tarefa de NER proporciona a distinção em frases ou texto entre o que é uma pessoa, uma organização, um lugar, uma data, dentre outras entidades definidas. Martins et al. (2020)

A maneira padrão de abordar o problema de NER é utilizando técnicas de rotulagem de sequências de palavras, onde as *tags* atribuídas capturam tanto o limite quanto o tipo de quaisquer entidades detectadas. De forma que seja possível diferenciar, por exemplo, os nomes próprios de um texto comum (Jurafsky and Martin, 2009). Em um texto padrão em português, uma sequência de palavras com a primeira letra em maiúsculo, ou com siglas como: "Dr.", "Sr.", dentre outros, antecedendo o texto, é suscetível ser um nome próprio. Por outro lado, se eles forem precedidos por palavras como "cidade", é possível que se trate de uma localização. Ou seja, para cada entidade é possível identificar sinais sobre fatos e contextos ao qual elas pertencem, e assim, definir um conjunto de características que permita identificar e classificá-las.

Nessa perspectiva, pode-se dizer que a tarefa de NER se parece muito com o problema da segmentação sintática e reconhecimento da classe gramatical. Na prática, nada mais é que realizar uma tarefa de classificação, na qual cada palavra contida em uma sentença será associada a um rótulo. Entretanto, essas palavras podem apresentar significados diferentes a depender de palavras adjacentes a elas, assim, a rotulagem de sequência para a identificação de entidades e suas características entra como uma alternativa para complementar a tarefa de classificação de uma forma mais assertiva. Na abordagem de rotulagem de sequência para NER, os classificadores são treinados para anotar os *tokens* em um texto com tags que indicam a presença de tipos particulares de entidades nomeadas (Jurafsky and Martin, 2009).

Alguns recursos padrões empregados no estado da arte de NER incluem o recurso de formação da palavra, o qual inclui a identificação de letras maiúsculas, minúsculas, bem como padrões mais elaborados, projetados para capturar expressões que usam números (A9), pontuação (Brasília!), dentre outros. Outro recurso utilizado é verificar a presença da entidade em um dicionário, por exemplo, o uso de listas contendo nomes de organização, pessoas e localizações para identificar entidades em um texto origem. Essa técnica é implementada como um vetor binário que atribui uma marcação para cada entidade identificada na lista (Jurafsky and Martin, 2009).

Considere a seguinte frase do exemplo da Figura 2.3.

Figura 2.3: Exemplo de uma frase com codificação *IOB*

A Figura 2.3 mostra a marcação de estilo *IOB* (*Inside, Outside, Beginning*), na qual se utiliza como padrão de captura palavra por palavra. Assim como no agrupamento sintático, o conjunto de *tags* para tal codificação consiste em uma ou mais *tags* para cada

tipo de entidade reconhecida, conforme pode ser observado na entidade *"Pessoa"* (*B-PER*, *I-PER*) e *"Localização"* (*B-LOC*, *I-LOC*), e "O" para qualquer entidade fora do interesse (*Outside*). Cada esquema de marcação é formado pela *tag* da marcação IOB, seguida pela *tag* do tipo de entidade, as quais permitem avaliar a qual classe a entidade pertence.

Segundo Sang and De Meulder (2003), utilizando essa abordagem é possível atribuir quatro categorias pré-definidas para as entidades: *"Pessoa"*, *"Organização"*, *"Localização"* e *"Miscelânea"*. A categoria *"Miscelânea"* é utilizada para as entidades que não pertencem a nenhuma das três primeiras.

Uma vez definida a codificação IOB, o próximo passo é definir as características associadas com cada entrada de texto, ou cada *token* que deverá ser anotado. Essas características devem ser preditores coerentes com o rótulo da classe e extraíveis de forma confiável do texto de origem e podem não ser baseados apenas nas características do *token*, mas também em características ao redor do *token* (Jurafsky and Martin, 2009).

Além da definição das entidades e do conjunto de dados, existem diferentes abordagens para subsidiar a ideia geral de como lidar com a tarefa de NER. O funcionamento mais detalhado da abordagem de aprendizado supervisionado fraco pode ser melhor compreendido na Seção 2.6.

Aplicações genéricas da tarefa de NER contemplam a detecção de entidades gerais conforme citado anteriormente. Entretanto, é possível criar aplicações especializadas que abrangem outros tipos de entidades como, por exemplo, *"número do contrato"*, *"CPF"*, *"E-mail"*, *etc.* (Jurafsky and Martin, 2009).

2.4 Modelos para a tarefa de Reconhecimento de Entidade Nomeada

Os modelos aplicados a tarefa NER podem ser divididos em duas categorias: os modelos tradicionais para NER e os modelos baseados em *Deep Learning* (DL). Na Seção 2.4.1 será apresentado o modelo Linear-chain Conditional Randon Fields (CRF), um modelo tradicional de sequência comumente utilizado para as tarefas de NER e na Seção 2.4.2 o modelo Bi-directional Long Short Term Memory (Bi-LSTM), um modelo para a tarefa de NER baseado nos princípios de DL.

2.4.1 Linear-chain Conditional Randon Fields (CRF)

O CRF é um algoritmo de modelagem de sequência que combina as vantagens da modelagem de classificação e da modelagem gráfica. O CRF possui a habilidade de modelar dados multivalorados de forma compacta com a capacidade de alavancar um número maior de características de entrada para realizar a previsão da classe. No geral, os CRF são uma junção de classe de modelos discriminativos e do modelo generativo, ao possuírem tanto a distribuição conjunta, quanto a distribuição condicional (Sutton and McCallum, 2012).

O CRF vem como uma proposta mais completa que o BoW. O BoW considera que a ausência ou a presença da palavra é mais importante do que a sequência de palavras, ignorando a posição exata das palavras (Jurafsky and Martin, 2009). No entanto, existem problemas de reconhecimento de entidade que a posição importa tanto, ou se não mais. Por exemplo, *New York City* e *New York Times*, a classificação observando a sequência

das palavras será completamente diferente nos dois casos. No primeiro, a classe será a de "Localização", enquanto a segunda a classe será de "Organização". Nesse sentido, o CRF considera que as características de cada palavra são dependentes uma da outra, além de realizar uma observação futura enquanto aprende o padrão de cada classe.

O CRF de forma probabilística realiza predição de rótulos de dados sequenciais considerando o contexto em que o dado está inserido. Uma implementação bastante comum do CRF é realizar a marcação NER combinada com a notação IOB (veja Seção 2.3), ao otimizar a predição das ocorrências entre as marcações IOB, tendo em vista que o modelo se limitará a identificar as ocorrências apenas das anotações possíveis no contexto. Dessa forma, o CRF apresenta bons resultados em termos de desempenho para o problema de NER (Sutton and McCallum, 2012)

Formulação do CRF

Uma abordagem tradicional para o problema de NER é organizar as variáveis de saída em uma sequência linear através do modelo de Markov, o HMM (Hidden Markov Model). No modelo HMM, para uma sequência de observações $X = \{x_t\}_{t=1}^{T}$ existe uma sequência de *estados* $Y = \{y_t\}_{t=1}^{T}$ pertencentes a um conjunto de estados finitos de S, ou seja, cada observação x_t é uma palavra na posição t e cada estado y_t é um rótulo da palavra. O HMM modela a distribuição conjunta $p(x, y)$ e faz duas suposições de independência dessas variáveis. Primeiro, ele assume que cada estado depende somente do estado predecessor imediato, isto é, dado um estado anterior y_{t-1}, cada estado y_t é independente de todos os antecessores $y_1, y_2, ..., y_{t-2}$. Segundo, ele também assume que cada palavra x_t depende somente do estado atual y_t.

Dessa forma, pode-se especificar um modelo HMM abrangendo três distribuições de probabilidade: primeiro, a distribuição $p(y_1)$ sobre os estados iniciais; segundo, a distribuição de transição $p(y_t|y_{t-1})$; e por último, a distribuição da observação $p(x_t|y_t)$. Resumindo, essas distribuições fornecem a probabilidade conjunta de uma sequência de estados y e uma sequência de observação x que pode ser descrita como:

$$p(y, x) = \prod_{t=1}^{T} p(y_t|y_{t-1})p(x_t|y_t),$$

onde o estado inicial da distribuição $p(y_1)$ foi descrito como $p(y_1|y_0)$

Embora haja vantagens nessa abordagem, ela possui limitações importantes, como, por exemplo, o tamanho da dimensão de x e a complexidade das dependências, portanto implementar uma distribuição de probabilidade somente se torna uma tarefa difícil, além de afetar o desempenho. Uma solução para este problema é modelar a distribuição condicional $p(y|x)$ diretamente. Uma solução para este problema é a implementação do CRF. O CRF é considerado uma distribuição condicional $p(y|x)$ que segue a mesma ideia da distribuição conjunta $p(y, x)$ de uma HMM . Porém, na abordagem do CRF o ponto principal da formulação é que a distribuição condicional é de fato um campo aleatório condicional com uma escolha particular de funções de característica semelhante ao que ocorre na regressão logística, que foca na distribuição condicional de $p(y|x)$. Ou seja, dado um vetor de características x, deseja-se prever um vetor de resultados $y = \{y_0, y_1, ..., y_r\}$ de um conjunto de variáveis aleatórias.

A implementação CRF como uma abordagem para NER assume que a variável y_s é uma *entidade* da palavra na posição s e o parâmetro de entrada x é dividido no vetor de características $x = \{x_0, x_1...x_T\}$, no qual cada x_s contém várias informações sobre a própria palavra, como, por exemplo, prefixos, sufixos, se começa com letra maiúscula, e *etc.*.

De acordo com Sutton and McCallum (2012), um CRF segue a seguinte Definição:

Definição 1 Seja Y, X vetores em R^K, $\theta = \{\theta_k\} \in R^K$ um parâmetro, e $\{f_k(y, y^{'}, \mathbf{x}_t)\}_{k=1}^{K}$ uma função correlacionando a sequência de características. Um CRF de cadeia linear é uma distribuição condicional de $p(y|x)$ que segue a seguinte equação:

$$p(y\|x) = \frac{1}{Z(x)} \prod_{t=1}^{T} exp \left\{ \sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, x_t) \right\}$$

Onde:

$Z(x)$ é a soma de todas as probabilidades possíveis para as características sobre cada sequências de estados possíveis com um número exponencial de termos.

θ_k é o peso de cada características indexada por k

K é a quantidade de características

$f_k(y_t, y_{t-1}, x_t)$ é a função que contém todas as informações sobre x necessárias para computar as características no estado t.

T é o número de estados de Y.

O problema de inferência no CRF busca resolver o problema $\arg\max_y p(y|x)$ que busca o maior y para maximizar a probabilidade de y dado x. Esse problema de inferência do CRF é intratável, porém existem vários algoritmos que retornam uma solução aproximada, conforme listado no artigo escrito por Mccallum Sutton and McCallum (2012).

2.4.2 *Bi-directional Long Short Term Memory*

Bi-LSTM em uma tradução livre para o português, significaria Memória de Curto e Longo prazos, e é um tipo de RNN (*Recurrent Neural Network*). Nas redes neurais tradicionais todas as entradas e saídas são independentes umas das outras. Porém, para o problema de NER predizer a próxima palavra de uma sentença é necessário que se conheça a palavra anterior, nesse caso a RNN guarda todas as informações da palavra anterior, pois a identificação do contexto da próxima palavra depende da palavra anterior (Schmidt, 2019). A RNN resolve este problema adicionando uma camada recorrente ao modelo das redes tradicionais conforme Figura 2.4

A RNN é uma rede amplamente utilizada para solucionar problemas sequenciais como, por exemplo, séries temporais ou mesmo textos, formados por um conjunto de palavras que, ordenadamente, geram sentido ao texto. Entretanto, sofrem com a dissipação do gradiente (Schmidt, 2019). Os gradientes são utilizados para atualizar os pesos da rede e reduzem consideravelmente, ao longo da propagação, chegando ao ponto de deixarem de contribuir para a atualização dos pesos.

Para solucionar esse problema, foram introduzidas novas características na RNN clássica para formar o LSTM, conhecidos como "portões". Um portão de entrada foi criado

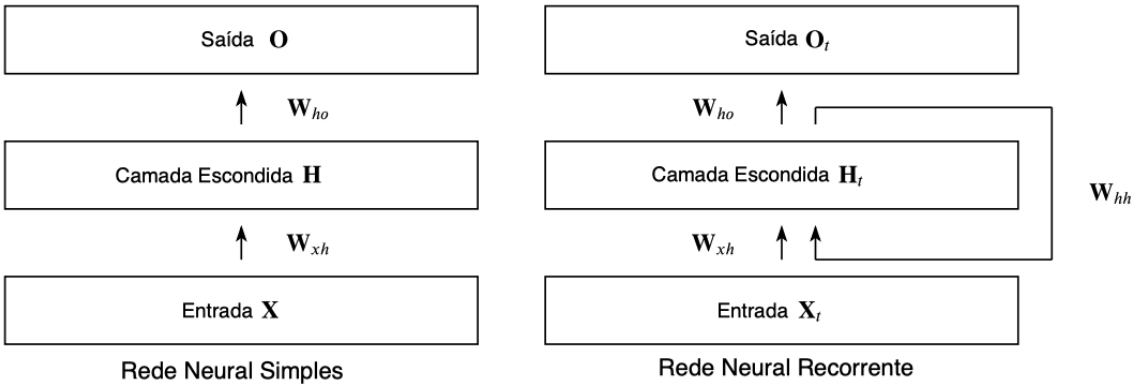


Figura 2.4: Rede Neural Recorrente RNN (Schmidt, 2019)

para proteger a memória armazenada. O portão de saída foi criado para proteger outras unidades de informações irrelevantes. Existe ainda a memória da célula, que recebe informações de estados passados e dos portões de entrada e saída (Hochreiter and Schmidhuber, 1997). A Figura 2.5 ilustra uma representação dos principais portões da arquitetura LSTM.

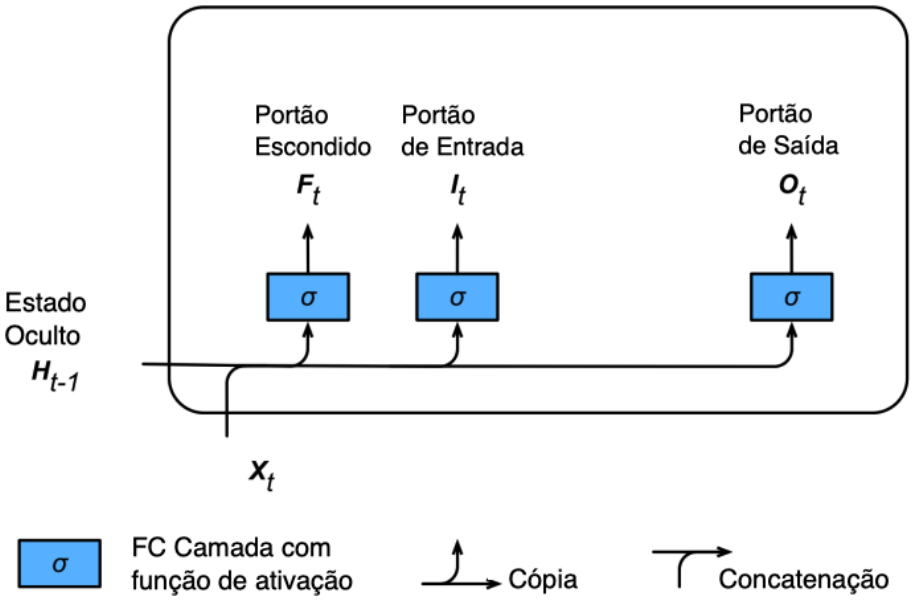


Figura 2.5: Arquitetura do LSTM (Schmidt, 2019)

O Bi-LSTM estende do LSTM, porém, introduz uma segunda camada escondida, na qual as conexões ocultas ocorrem em ordem temporal oposta a outra camada escondida. Ou seja, o modelo captura tanto informações do passado como do futuro. O componente CNN é utilizado para induzir as características ao nível de caracteres.

A Figura 2.6 ilustra a estrutura da arquitetura do CNN Bi-LSTM.

As características extraídas de cada palavra são alimentadas em uma rede LSTM para frente e uma rede LSTM para trás. A saída de cada rede em cada passo de tempo é decodificado por uma camada linear e uma camada de probabilidades e uma camada

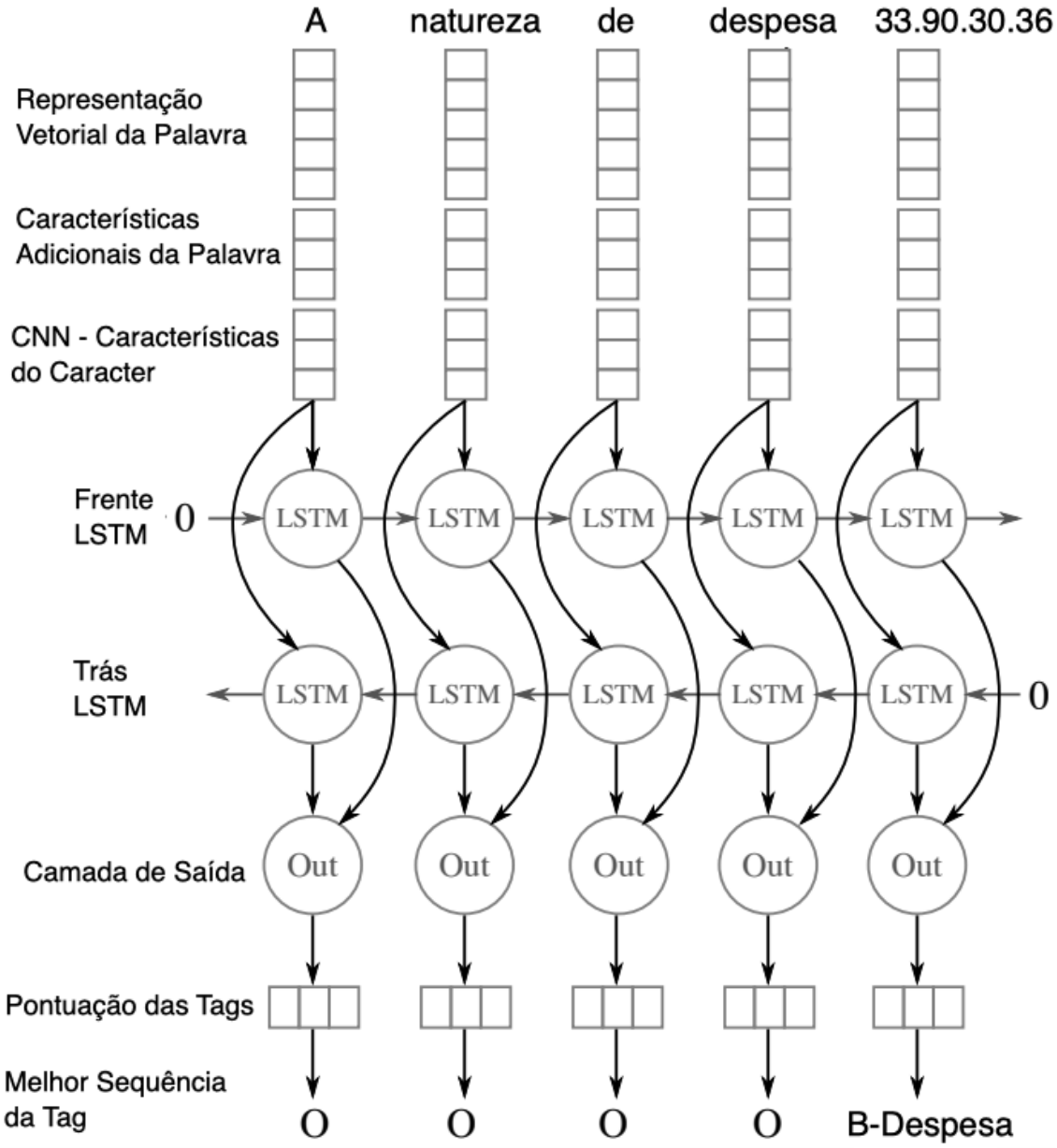


Figura 2.6: Arquitetura do CNN Bi-LSTM (Chiu and Nichols, 2015)

para cada categoria de *tag*. Um novo vetor extrai as características por caractere, como incorporações de caracteres e (opcionalmente) tipo de caractere. Esses vetores são então somados para produzir a melhor sequência de *tag*.

2.5 Aprendizado de Máquina

O aprendizado de máquina é definido como o paradigma computacional onde a capacidade de resolver o problema dado é construída com base em exemplos anteriores. A ideia básica do aprendizado de máquina é o raciocínio baseado em exemplos, que é o processo de raciocinar sobre o problema com base dos exemplos anteriores semelhantes ou exemplos

de treinamento. Ou seja, realiza-se o aprendizado sobre casos semelhantes para resolver os problemas reais a partir da generalização (Jo, 2021).

O aprendizado de máquina pode ocorrer de várias formas, como, por exemplo, o aprendizado supervisionado, não supervisionado, semi-supervisionado e por reforço.

2.5.1 Aprendizado Supervisionado

Neste tipo de aprendizado, o modelo é treinado com um conjunto de dados rotulados, ou seja, dados que já possuem a resposta correta para o problema que se deseja resolver. O modelo aprende a mapear os dados de entrada para a saída desejada, utilizando algoritmos de regressão e classificação, como regressão linear, regressão logística, k-Nearest Neighbors (KNN), redes neurais artificiais (ANNs), entre outros.

No aprendizado supervisionado, é necessário associar cada exemplo de treinamento com seu próprio rótulo. Durante o processo de aprendizado, os parâmetros são otimizados para minimizar a taxa de classificação incorreta ou o erro (Jo, 2021).

Como exemplo desse aprendizado pode-se citar: classificação de e-mails como spam ou legítimo, previsão de preços de ações e detecção de fraudes

2.5.2 Aprendizado Não Supervisionado

Neste tipo de aprendizado, o modelo é treinado com um conjunto de dados não rotulados, ou seja, dados que não possuem a resposta correta. O modelo aprende a encontrar padrões e estruturas nos dados sem a necessidade de rótulos pré-definidos. O aprendizado não supervisionado contempla o processo de otimizar os protótipos dos clusters, com base nas similaridades entre os exemplos de treinamento. Algoritmos como k-means clustering, análise de componentes principais (PCA), e t-SNE são comumente utilizados. Como exemplo desse aprendizado pode-se citar: agrupar clientes em diferentes segmentos, segmentar imagens e descobrir tópicos em documentos (Jo, 2021).

2.5.3 Aprendizado por Reforço

Neste tipo de aprendizado, o modelo aprende a tomar decisões através da interação com um ambiente. O modelo recebe recompensas ou penalidades por suas ações, e aprende a tomar as melhores decisões para maximizar sua recompensa ao longo do tempo. Algoritmos como Q-learning e Monte Carlo Tree Search (MCTS) são comumente utilizados. Como exemplo desse aprendizado pode-se citar: treinar robôs para realizar tarefas complexas, jogar jogos e Otimizar sistemas de controle (Jo, 2021).

2.5.4 Aprendizado Semi-Supervisionado

Combina os princípios do aprendizado supervisionado e não supervisionado. Por exemplo, quando casos de treinamento não rotulados são utilizados com os rotulados para o processo de aprendizado. O modelo aprende com exemplos rotulados minimizando o erro entre os rótulos previstos e os reais. Ele aprende com exemplos não rotulados aproveitando suas similaridades com os dados rotulados. Os exemplos não rotulados são agrupados usando um algoritmo de aprendizado não supervisionado, que considera os exemplos rotulados.

O algoritmo de aprendizado supervisionado é treinado com ambos os tipos de exemplos. Uma maneira comum de implementar o aprendizado semi-supervisionado usando essa combinação é utilizar o Naive Bayes e o algoritmo EM (Jo, 2021).

2.6 Aprendizado Fracamente Supervisionado

Em cenários normais, os dados utilizados para treinar modelos de aprendizado de máquina são anotados de forma manual por especialistas. Entretanto, obter estas coleções de dados anotados nem sempre é viável em termos de tempo e custo. Assim, a necessidade de aplicar NER a contextos no qual não há dados anotados (Sugiyama et al., 2022) impõe a investigação de outras abordagens.

Os dados não rotulados originalmente podem ser "rotulados artificialmente"com base em similaridades ou em classificadores que os aprendem. A adoção de técnicas como o aprendizado ativo e por transferência visam minimizar a necessidade de dados anotados. O aprendizado ativo trabalha com a suposição de que um grande conjunto de dados pode ser bem representado por um subconjunto menor de suas amostras e, idealmente, um modelo treinado com o conjunto de dados completo ou com esse subconjunto representativo menor obtém desempenho semelhante. Enquanto no aprendizado por transferência uma variante bidirecional relaciona várias tarefas resolvidas simultaneamente pelo compartilhamento mútuo de informações (Sugiyama et al., 2022). Essas técnicas objetivam melhorar o processo de coleta e anotação dos dados, buscando ampliar a quantidade de dados para treinamento. No entanto, elas ainda requerem alguma quantidade de dados anotados (Lison et al., 2020).

O uso das técnicas de aprendizado fracamente supervisionados, em especial a abordagem de SF combinada com a tarefa de NER surgem como uma alternativa para o problema de escassez de dados anotados (Lison et al., 2020). Esta abordagem baseia-se em uma ampla coleção de técnicas de aprendizagem de máquina para realizar indicações de rótulos e assim treinar um modelo com base em dados que contêm informações “mais fracas”, do que dados totalmente supervisionados (Tok et al., 2022) (Sugiyama et al., 2022).

Os rótulos formados por dados que contêm informações “mais fracas”, do que dados totalmente supervisionados são considerados rótulos ruidosos ou rótulos fracos. Embora os termos "rótulos ruidosos"e "rótulos fracos"sejam frequentemente usados de forma intercambiável, eles podem ter nuances diferentes. Os rótulos ruidosos referem-se a etiquetas incorretas, imprecisas ou inconsistentes atribuídas aos exemplos de dados, seja por erro humano na anotação, imprecisão nos métodos de coleta de dados ou ambiguidades nos critérios de anotação. Enquanto os rótulos fracos são etiquetas que, embora possam estar corretas, fornecem informações limitadas ou menos confiáveis sobre os exemplos de dados, seja porque as informações disponíveis para rotular os dados são insuficientes, vagas ou indiretas, resultando em uma associação menos precisa entre os rótulos e os verdadeiros conceitos subjacentes.

Para Tok et al. (2022) existem três tipos de SF: supervisão incompleta, onde apenas um subconjunto de dados de treinamento de alta qualidade é anotado; supervisão inexata, onde apenas rótulos não completos ou correlatos são dados; supervisão imprecisa, onde somente rótulos ruidosos ou fracos são informados.

A supervisão imprecisa diz respeito à situação em que as informações de supervisão nem sempre são verdadeiras; em outras palavras, algumas informações de rótulo podem

sofrer de erros. Entre as várias abordagens e características da SF imprecisa, destaca-se a possibilidade de anotar dados de forma programática com o uso de coleções de FR que automatizam e anotam cuidadosamente os documentos com rótulos de entidades nomeadas (Lison et al., 2020)(Ratner et al., 2017)(Safranchik et al., 2020). Mais formalmente, Tok et al. (2022) define um novo paradigma em que é possível gerar um grande conjunto de dados treinados de forma programática. A programação dos dados, como é denominada, orienta a criação de várias FR de uma forma organizada, as quais são utilizadas para anotar cada exemplo identificado.

Um típico exemplo dessa proposta, apresentado por Ratner et al. (2017), é o *framework* chamado *Snorkel* [4]. Esse *framework* permite aos usuários treinar modelos de aprendizado de máquina apenas escrevendo funções que combinam várias fontes de supervisão e usando um modelo generativo para estimar a precisão (possíveis correlações) de cada fonte (Bach et al., 2019).

Além da geração das FR, a programação do dado sugere a utilização de outros mecanismos que possibilitam reduzir a geração de rótulos ruidosos, como, por exemplo, a adoção de um modelo generativo (Tok et al., 2022). Nesse sentido, o trabalho de Lison et al. (2020) implementou o Skweak [5] uma arquitetura que permite aplicar mais de uma função de rótulo às entidades e unificar as FR ruidosas em uma única anotação, considerando a precisão de cada função de rótulo.

A Figura 2.7 apresenta uma visão geral da arquitetura de um *framework* que implementa a abordagem sugerida pela Programação do Dado para SF:

- passo 1 representa as FR a serem aplicadas;
- passo 2 exibe a adoção do modelo generativo, visando realizar a agregação dos rótulos;
- passo 3 treina o modelo a partir dos rótulos gerados para cada exemplo.

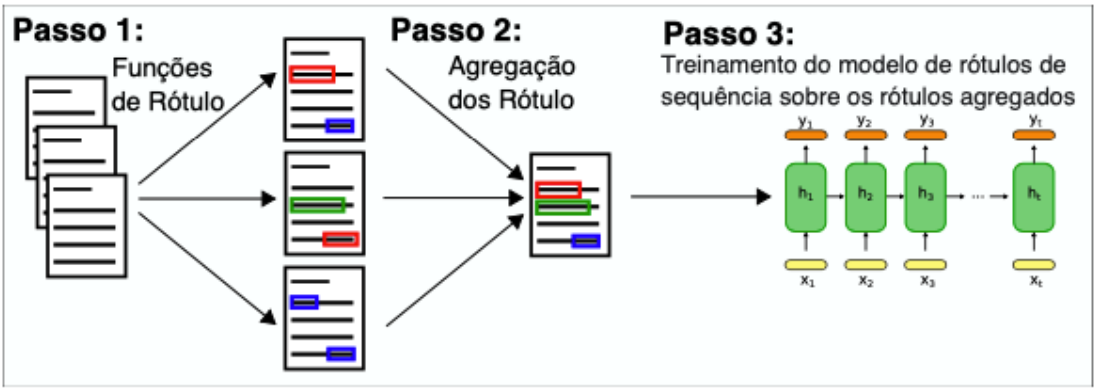


Figura 2.7: Visão Geral da Arquitetura da Programação do Dado para SF (Lison et al., 2020)

A abordagem da SF pode incluir técnicas como aprendizado semi-supervisionado, aprendizado ativo, aprendizado por transferência e aprendizado autos supervisionado,

[4] `https://snorkel.ai/`
[5] `https://github.com/NorskRegnesentral/skweak`

entre outras, que permitem ao modelo aprender com uma quantidade menor ou menos precisa de rótulos, aproveitando outras fontes de informação disponíveis nos dados (Bishop, 2006). Nesse estudo, utilizou-se a supervisão fraca por meio da programação de dados que adota o uso de FR, as quais geralmente se baseiam em técnicas de aprendizado semi-supervisionado. No aprendizado semi-supervisionado, o modelo é treinado com uma combinação de dados rotulados e não rotulados. As FR são usadas para atribuir rótulos aos dados não rotulados, o que permite que o modelo aprenda com uma quantidade maior de exemplos durante o treinamento (Tok et al., 2022).

2.6.1 Funções de Rótulos

As funções de rótulos são essencialmente regras codificadas que os engenheiros de dados ou especialistas de domínio usam para atribuir cada amostra de dados a uma classe específica. As FR podem ser escritas de várias maneiras, mas a melhor maneira de começar a criá-las é mapear as características dos subconjuntos de dados (Tok et al., 2022).

As FR fornecem rótulos fracos potencialmente ruidosos que alimentam todo o *pipeline* de aprendizagem e podem ser desenvolvidas utilizando diferentes fontes de aprendizagem fracamente supervisionado (Tok et al., 2022). A Definição 2 descreve formalmente a definição de FR.

Definição 2 Função de Rótulo é uma função qualquer $\phi : X \rightarrow Y \cup \{-1\}$ que recebe como entrada um ponto de dados em X e as saídas são rótulos em Y ou a abstenção de rótulo representado como -1.

Fundamentalmente, as FR não precisam fornecer uma previsão para cada ponto de dados e pode "abster-se" sempre que certas condições não forem satisfeitas. Estas condições podem ser definidas com as mais diversas técnicas de aprendizagem fracamente supervisionado. Para efeito desta pesquisa, as FR serão agrupadas em 3 categorias: baseadas em correspondência por palavra, conhecimento heurístico (regras) e aprendizado de máquina.

Função de Rótulos Baseada em Aprendizado de Máquina

As FR do tipo aprendizado de máquina podem assumir a forma de modelos de aprendizado treinados em dados de outro domínio relacionado, levando assim alguma forma de transferência de aprendizagem entre os domínios (Lison et al., 2021). Por exemplo, os rótulos *GPE* (*Geo-Political Entity*) e *LOC* (*Location Name*) na base de dados Ontonotes (Weischedel, 2013), de granulação mais específica, podem ser utilizados como rótulos de referência para o rótulo *LOC* de granulação mais geral na base de dados CoNLL 2003 (Sang and De Meulder, 2003). Por exemplo, Ratner et al. (2017) utiliza modelos pré-treinados na base *DBPedia*(Lehmann et al., 2015) para procurar relacionamentos conjugais em notícias e Safranchik et al. (2020) aproveita o modelo de tópico semântico existente para identificar conteúdos irrelevantes para a categoria de produtos de interesse.

Um exemplo dessa função pode ser observada no algoritmo 1 o qual realiza a rotulação de *tokens* de texto para a entidade do tipo "Pessoa". Neste exemplo foi utilizado um

modelo de aprendizado de máquina já treinado na base de dados multi-língua e padrão prata *WikiNER* (Nothman et al. (2013)) e disponibilizado pela biblioteca *Spacy*[6].

Algorithm 1 Exemplo de Função de Rótulo do Tipo Aprendizado de Máquina

```
1: procedure PESSOA(string texto)
2:     nlp_model = spacy.load("pt_core_web_sm")
3:     for token in texto do doc = nlp_model(token)
4:         if tag = "Pessoa" then
5:             return token.i-1, token.i+1, "Pessoa"
6:         end if
7:     end for
8: end procedure
```

Função de Rótulos do Tipo Correspondência por Palavra

As FR inspiradas em correspondência buscam a ocorrência da entidade em uma lista de palavras ou frases. Essas funções são úteis para identificar se uma determinada entidade está presente em um conjunto de dados. Elas podem ser implementadas de várias maneiras, como por meio de correspondência exata ou correspondência parcial. O objetivo é encontrar correspondências entre a entidade e as palavras ou frases da lista, permitindo assim a categorização ou anotação adequada dos dados.

Por exemplo, a partir de um dicionário de nomes, como a base de nomes do IBGE [7] realizar a anotação de entidades do tipo *"PESSOA"*. Outro exemplo, é o uso de uma base geográfica, como a Geonames (Wick, 2015) realizar a anotação de entidades do tipo *"LOC"*, conforme função descrita no Algoritmo 2.

Algorithm 2 Exemplo de uma função do tipo Correspondência por Palavra

```
1: procedure PESSOA(string texto)
2:     nomes= [("Lucelia", "Mota"), ("Thiago", "Faleiros")]
3:     for token in texto do
4:         if token in nomes then
5:             return token.i-1, token.i+1, "Pessoa"
6:         end if
7:     end for
8: end procedure
```

Nesta pesquisa aplicou-se o método da correspondência exata, atribuindo um rótulo as entidades que eram encontradas na lista de valores definidos para a categoria

Função de rótulos do Tipo Conhecimento Heurístico (regras)

Em aplicações práticas, os usuários geralmente têm conhecimento de domínio sobre a tarefa de aprendizagem de seu interesse. Um tipo comum de FR é expressar o conhecimento

[6] `https://spacy.io/`

[7] Disponível em `https://basedosdados.org/dataset/br-ibge-nomes-brasil/`

do domínio em regras que associam os rótulos correspondentes aos conjuntos de dados. Por exemplo, em aplicativos de texto, os engenheiros de dados escrevem FR baseados em palavra-chave ou regras que atribuem rótulos correspondentes aos *tokens* de dados que contêm a palavra-chave ou corresponde à expressão regular especificada (Ratner et al., 2017).

Como exemplo do tipo de função de rótulos baseadas em conhecimento heurístico (regras), suponha uma entidade do tipo *"ÓRGÃO"*. Essa função baseia-se em uma regra para procurar trechos de texto que terminem com um tipo de órgão legal (como "Inc."). Da mesma forma, implementar regras que permitam identificar as entidades do tipo *"MOEDA"*, ao procurar por números precedidos por um símbolo de moeda (como o símbolo $). Veja o exemplo da função descrito no Algoritmo 3.

Algorithm 3 Exemplo de uma função do tipo conhecimento heurístico (Regras)

```
1: procedure MOEDA(string texto)
2:     for token in texto do
3:         if token[0].isdigit() an token(-1).iscurrency then
4:             return token.i-1, token.i+1,"Moeda"
5:         end if
6:     end for
7: end procedure
```

2.6.2 Modelos Generativos

As FR são geralmente ruidosas, o que significa que elas possuem alta taxa de erro. Após serem aplicadas a um conjunto de texto, as saídas das FR são agrupadas utilizando o modelo generativo (Lison et al., 2021). Nguyen et al. (2017), propõem um HMM para agregar rótulos dispersos e sequenciar anotações com objetivo de prover melhorias para a tarefa de NER. Segundo Rabiner and Juang (1986) a ideia principal do HMM é observar uma sequência de símbolos discretos ou contínuos, e propor um modelo que explica e caracteriza a ocorrência do símbolo observado.

A Figura 2.8 ilustra a arquitetura definida em Lison et al. (2020), a qual prevê uma abordagem que se baseia em FR que anotam documentos automaticamente com rótulos de entidade nomeada. Um módulo de agregação, baseado no modelo HMM, é utilizado para unificar as saídas das FR em uma única anotação (probabilística) conforme a precisão e revocação de cada função (Lison et al., 2020).

A adoção da abordagem de aprendizagem fracamente supervisionado pode apresentar rótulos com alto nível de ruído, por exemplo, alguns rótulos gerados podem não ser precisos e apresentarem um nível de erro. Entretanto, para Tok et al. (2022) e Lison et al. (2020), uma abordagem combinada do uso de uma base dados de rótulos "forte" tradicionais em conjunto com uso de rótulos "fraco" pode resultar em um modelo que aprende bem e com um bom desempenho.

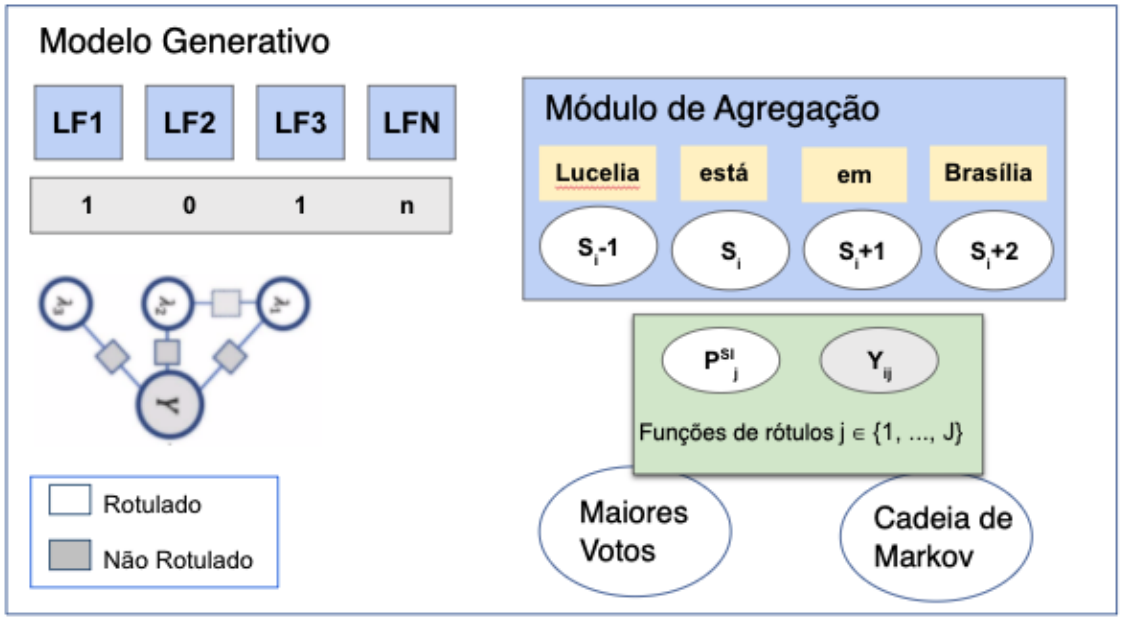


Figura 2.8: Visão Geral do Modelo Generativo (Lison et al., 2020)

Capítulo 3

Revisão da Literatura

A Revisão da literatura é um método para pesquisa bibliográfica realizada a partir de princípios prescritos na Teoria do Enfoque Meta Analítico Consolidada (TEMAC) Mariano (2017). O TEMAC sugere na etapa de preparação da pesquisa a construção da palavra-chave da pesquisa e a delimitação das fronteiras a serem pesquisadas. Neste capítulo será apresentado os resultados do protocolo TEMAC, o protocolo pode ser visualizado no Anexo A .

Para a construção da palavra-chave de pesquisa foram utilizadas palavras em inglês, envolvendo expressões relacionadas a SF e tarefa de NER, sendo assim foi formado o seguinte termo:

> *"weak supervision" or "weakly Supervised" or "weakly supervise") AND ("Named Entity" or "Tagging Sequence" or "Sequence Labelling" or "Labelling expression".*

Foram escolhidas para a pesquisa as bases de dados *Web of Science* e *Scopus* por serem base de dados consolidadas e de maior qualidade. Na *Web of Science*, foi aplicado o filtro pela categoria *Computer Science Artificial Intelligence* e na *Scopus*, pela categoria *Computer Science* com objetivo de tornar o resultado da pesquisa mais refinado. Por fim, foram retornados 24 artigos na *Web of Science* e 54 na *Scopus*. Esses trabalhos incluem utilizar técnicas que buscavam resolver o problema de NER, além de abordar diversas propostas de arquiteturas para aplicação da abordagem da SF, para criação de base de dados anotados. Dentre os artigos retornados, foram selecionas os que se mostraram pertinentes, pois exploravam a tarefa de NER combinada com abordagens de geração de rótulos de forma automatizada, além de iniciativas brasileiras para criação de bases anotadas.

3.1 Trabalhos correlatos

Nesta seção serão apresentados os trabalho que exploram a tarefa de NER e abordagem da SF em geral, alguns trabalhos que implementam a SF com aplicação de FR e por fim, algumas iniciativas brasileiras para criação de bases anotadas.

Em Li et al. (2021) os autores estudaram o problema de construção de sistemas de rotulação de entidades usando algumas regras lógicas para treinar um modelo neural automatizadamente. O trabalho de Li et al. (2021) propõe o *framework* TALLOR *(Tagging*

with Learnable Logical Rules), o qual visa incorporar regras lógicas que podem ser facilmente compreendidas e interpretadas por seres humanos. Isso ajuda a aumentar a confiança nas decisões tomadas pelo modelo e também a identificar possíveis padrões ou relações subjacentes nos dados. Dessa forma é gerado um conjunto de entidades candidatas e regras candidatas as quais são aplicadas em um *dataset* não anotado. Em seguida, é treinado um modelo neural com os exemplos de treinamento selecionados e os rótulos preditos. Por último, selecionam-se novas regras lógicas de regras candidatas usando as previsões. Essas novas regras aprendidas serão usadas na próxima iteração para obter rótulos de treinamento.

Em Safranchik et al. (2020) é proposto um *framework* para treinamento de modelos com SF por meio do uso de múltiplas regras baseada em heurística, além de oferecer um conjunto de regras de suporte que atribuem votos as *tags* de saída, nesse estudo foi proposto o uso de um novo tipo de mecanismo de SF, as chamadas regras de vinculação, que votam em como os elementos da sequência devem ser agrupados em *spans* com a mesma *tag*. Para estimar as precisões das regras e combinar suas saídas conflitantes em dados de treinamento é utilizado um modelo generativo baseado no modelo HMM.

O trabalho de Shang et al. (2018) propõe o *AutoNER*, o qual explora dois modelos neurais. O primeiro baseado na estrutura de rotulação de sequência tradicional com uma camada *Fuzzy-LSTM-CRF* modificada para lidar com *tokens* com múltiplos rótulos, porém com geração de rótulos ruidosos. O segundo, foi utilizado para lidar com os rótulos ruidosos por meio de um modelo neural adaptado para ser mais eficaz por meio da incorporação de frases de alta qualidade para reduzir os rótulos falsos negativos. Especificamente, para cada dois *tokens* adjacentes, a conexão entre eles é anotado como "Correspondente", quando os dois *tokens* são combinados com a mesma entidade; "Desconhecido", se pelo menos um dos *tokens* pertencer a uma palavra com rótulo de tipo desconhecido; e por último, "Quebra", caso não corresponda a nenhuma palavra anotada. A proposta é utilizar um *corpus* e um dicionário para gerar entidades anotadas por combinação exata de caracteres, e onde houver conflito na geração dos rótulos, os conflitos são resolvidas maximizando o número total de *tokens* correspondentes.

No trabalho de Lison et al. (2020) é possível observar a utilização de um amplo conjunto de FR, as quais são utilizadas para anotar textos sem rótulos de forma automática. A proposta sugere a indicação de anotações para uma entidade específica, que pode levar à geração de rótulos com ruído. Dessa forma, pode ocorrer mais de uma proposta de anotação para uma determinada entidade. Com objetivo de resolver esse problema, um modelo oculto de *Markov* HMM é treinado para unificar as anotações, em uma única anotação probabilística, considerando as métricas de precisão e cobertura de cada FR. Por fim, um modelo pode finalmente ser treinado a partir da base gerada por estas anotações.

O trabalho de Knofczynski et al. (2022) propõe a aplicação de um conjunto de FR para programaticamente gerar rótulos para os dados trafegando sobre a Internet. A proposta é aplicar essas funções sobre uma base de dados não rotuladas e gerar uma base de dados com rótulos fracos. A partir da base gerada pela supervisão fraca é possível realizar o treinamento por meio de múltiplos classificadores, em detrimento de apenas um. Os resultados demonstraram que o aprendizado sobre *datasets* menores utilizando múltiplas tarefas foi 2,3% mais acurado e 96% mais rápido que com uma única tarefa. Enquanto com *datasets* maiores a acurácia saltou para 41,2% e 94,1% mais rápido. O estudo concluiu que o uso da abordagem de supervisão fraca combinada com o aprendizado de múltiplas tarefas

podem ser aplicados em outros campos da ciência com níveis de resultados semelhantes ao alcançado.

Em Chen and Ding (2022) é proposto o uso de múltiplas abordagens de supervisão fraca, que podem anotar dados não rotulados automaticamente. Este trabalho se diferencia pelo fato de propor um método que gera os rótulos de forma probabilística por meio de um modelo condicional independente. A proposta reúne especificamente o uso de três tipos de supervisão fraca: dicionário de palavras, expressões regulares e supervisão distante por clusterização. As duas primeiras foram utilizadas para conhecimento explícito, e a última para conhecimentos tácitos. A supervisão distante por clusterização, assume que rótulos fracos podem ser obtidos baseados na indicação de um cluster, ou seja, por meio da suposição do cluster pode-se entender que o dado de entrada ao possuir a mesma estrutura de cluster pertence a mesma categoria, conforme preconiza a hipótese de clusterização do aprendizado semi-supervisionado. Os resultados alcançados demonstram que o uso de múltiplas abordagem de anotação por supervisão fraca alcança um *F1 Score* de 86% contra 78% da anotação por supervisão fraca simples.

No Brasil, a quantidade de documentos legislativos e jurídicos produzidos na última década aumentaram dramaticamente, tornando difícil para os profissionais do direito consultar e atualizar a legislação. Os sistemas de NER têm o potencial inexplorado para extrair informações de documentos oficiais, o que pode melhorar a recuperação de informações e os processos de tomada de decisão, nesse sentido alguns pesquisadores iniciaram estudos para explorar o domínio jurídico e legislativo. O trabalho de Luz de Araujo et al. (2018) se concentrou na tarefa de anotação de sentenças judiciais de granularidade fina realizada por estudantes de direito. As entidades foram mapeadas em um estudo preliminar composto por quatro entidades nomeadas jurídicas mais amplas e vinte e quatro entidades aninhadas (de granularidade fina). O *corpus* fornecido contém trechos de 594 decisões (62.933 frases; 1.782.395 *tokens*; 33.055 anotações de granularidade mais ampla e 57.573 anotações de granularidade fina). Os melhores resultados foram obtidos com o algoritmo CRF em que a pontuação *F1 Score* excedeu 90% para a maioria das entidades nomeadas.

O trabalho de Albuquerque et al. (2022) propôs a criação de um *corpus* de Documentos Legislativos Brasileiros para NER com referenciais de qualidade. O *corpus* criado era composto por projetos de lei e consultas legislativas da Câmara dos Deputados do Brasil. Os modelos CRF e HMM resultaram em um promissor *F1 Score* de 80,8% na análise por categorias e 81,04% na análise por tipos. As entidades com os melhores resultados médios no *F1 Score* foram "FUNDlei" e *"DATA"*, e as com os piores resultados foram *"EVENTO"* e *"PESSOA"*. O *corpus* também foi avaliado usando as arquiteturas Bi-LSTM CRF e *Glove*, alcançando *F1 Score* de 76,89% na análise por categorias e 59,67% na análise por tipos.

A revisão da bibliografia abordou diversos trabalhos relevantes no campo da supervisão fraca para a tarefa de anotação de dados. Shang et al. (2018), Safranchik et al. (2020), e Li et al. (2021) concentraram-se na identificação de regras e na construção de funções de correspondência e regras para gerar rótulos automaticamente. Enquanto isso, Lison et al. (2020) explorou a construção de funções de dicionário, regras e modelos de aprendizado de máquina pré-treinados, com uma camada de agregação para indicar os rótulos fracos com maior confiança. Os projetos Knofczynski et al. (2022) e Chen and Ding (2022) desenvolveram frameworks que realizavam rotulação fraca com diferentes abordagens e

compararam os resultados. Em Safranchik et al. (2020), Li et al. (2021) e Lison et al. (2020), uma camada de agregação foi implementada após a geração das FR, enquanto nos demais trabalhos essa técnica não foi identificada.

Todos os trabalhos produziram um corpus por meio do treinamento com os rótulos indicados pela supervisão fraca. No entanto, as iniciativas brasileiras, como Luz de Araujo et al. (2018) e Albuquerque et al. (2022), limitaram-se ao treinamento do modelo em bases anotadas manualmente.

Diante desse panorama, surgiu a oportunidade de explorar a aplicabilidade dessas técnicas no contexto da língua portuguesa e da Licitação e Contratação Pública. Assim, esta pesquisa desenvolveu um conjunto de funções de rótulo conforme proposto em trabalhos anteriores e adotou a camada de agregação sugerida por Lison et al. (2020) para criar um corpus brasileiro de Licitações Públicas. Essa abordagem foi semelhante aos trabalhos de Luz de Araujo et al. (2018) e Albuquerque et al. (2022), mas sem depender da existência de um conjunto de dados rotulados manualmente. Como resultado, foi possível treinar um modelo com base no corpus gerado, que pode ser utilizado para rotular dados relacionados ao contexto de Licitação e Contratações Públicas em língua portuguesa.

3.2 Análise Comparativa

A tabela 3.1 apresenta uma comparação dos trabalhos correlatos. A coluna *Corpus* indica os conjuntos de dados de Licitação e Contratação Pública (CP) utilizados para validar os resultados. A coluna *Dados* fornece informações detalhadas sobre esses conjuntos de dados, enquanto a coluna *Rótulo* descreve os tipos de rótulos gerados. A coluna *Métodos* lista os métodos empregados para o treinamento dos modelos, e a coluna *F1 Score* mostra os resultados alcançados em cada trabalho. Na coluna *Características*, são destacadas as principais características dos trabalhos, identificadas por meio de uma análise interpretativa dos artigos. Por fim, a coluna *Diferencial* destaca o que esta pesquisa inova em relação aos trabalhos avaliados.

Tabela 3.1: Comparação dos trabalhos correlatos.

Referência	Corpus	Dados	Rótulos	Métodos	F1 Score	Características	Diferencial
(Li et al., 2021)	BC5CDR, CHEMD-NER, CoNNL 2003	BC5CDR (15.953 medicação e 13.318 entidades de doenças), CHEMDNER (10.000 Medicações e 84.355 doenças) e CoNLL2003 (14.041 train, 3.250 dev e 3.453 test)	Doenças e Medicação	TALLOR	BC5CDR (66.73%), CoNLL2003 (64.22%) e CHEMDNER (61.56%)	No trabalho de Li et al. (2021) as regras lógicas são aprendidas automaticamente, a partir de exemplos de treinamento que contêm apenas informações parciais sobre as entidades nomeadas a serem marcadas. Como ponto de atenção temos a questão de as regras identificadas poderem não distinguir dois conceitos semânticos intimamente relacionados, apresentando erros para identificar as fronteiras das entidades e para identificar múltiplas entidades como, por exemplo, *"HIT tipo II"* e sua sub-categoria *"HIT"*, rotulando apenas umas das entidades com rótulo ouro.	Esta pesquisa, apesar de utilizar FR do tipo regras, não utilizou exemplos parcialmente anotados, mas sim o conhecimento heurístico (critérios pré-definidos) para desenvolver as regras que capturavam as características da entidade. Outro diferencial, foi adoção de outros mecanismos para identificação do rótulo, como a correspondência por palavras e modelos treinados por aprendizado de máquina.
(Safranchik et al., 2020)	BC5CDR, NCBI-Disease, LaptopReview	BC5CDR (20.217), NCBI-Disease (7.286 doenças) e LaptopReview (3.845 sentenças)	Doenças, Medicação e Termos Técnicos	CRF e Bi-LSTM	BC5CDR (83.28%), NCBI-Disease (79.03%) e LaptopReview (69.04%)	O trabalho de Safranchik et al. (2020) utilizou o conhecimento heurístico (critérios predefinidos) para guiar o processo de rotulação. Entretanto, as regras têm precisão desconhecida e o processo de resolução dos conflitos por meio de votos não é transparente. Muitas vezes é natural que os usuários tenham inferências diferentes sobre qual *tag* um elemento deve ter e quão longe essa decisão deve se propagar para *tags* vizinhas. Por exemplo, o *token* "Inc." provavelmente faz parte de uma organização. Outra dificuldade é o mapeamento de várias dessas regras porque existem *tags* que ainda não foram inferidas.	Esta pesquisa utilizou a abordagem do conhecimento heurístico (critérios pré-definidos) para construção das funções de regras, mas acrescentou outros mecanismos para identificação do rótulo, como a correspondência por palavras e modelos treinados por aprendizado de máquina, a fim preencher a lacuna do trabalho anterior de identificação dos rótulos não mapeados pelas regras. Além de adotar uma camada com o modelo de agregação para definição do rótulo final.
(Shang et al., 2018)	BC5CDR, NCBI-Disease, LaptopReview	BC5CDR (15.953 medicações e 13.318 doenças), NCBI-Disease (6.866 doenças) e LaptopReview (3.845 sentenças)	Doenças, Medicação e Termos Técnicos	*FUZZY* LSTM CRF e AutoNER	BC5CDR (84.8%), NCBI-Disease (75.52%), LaptopReview (65.44%)	O trabalho de Shang et al. (2018) utilizou dicionários que contém termos relevantes e específicos para guiar o processo de rotulação. Entretanto, o dicionário pode não conter todas as entidades do domínio, além de novas entidades surgirem o tempo todo e dicionário ter que estar sempre sendo atualizado	Esta pesquisa utilizou a abordagem da correspondência de palavras, semelhante ao mecanismo de dicionários, mas a fim de preencher a lacuna do trabalho anterior, acrescentaram-se FR de conhecimento heurístico (regras) e modelos treinados por aprendizado de máquina. Além de adotar uma camada com o modelo de agregação para definição do rótulo final.
(Lison et al., 2020)	CoNNL Reuters and Bloomberg 2003	1163 documentos e 35.089 entidades 1.054 sentenças	*"ORG"*, *"PER"*, *'LOC"*, *"MISC"* *'PERSON"*, *"NORP"*, *"ORG"*, *"LOC"*, *"PRODUT"*, *"DATETIME"*, *"PERCENT"*, *"MONEY"*, *"QUANTITY'*	HMM-aggregatedlabels Neural HMM-agg	CoNNL (70.2% a 71.6%) Reuters and Bloomberg (68% a 81.6%)	O trabalho de Lison et al. (2020) utilizou funções de conhecimento heurístico (regras), correspondência de palavras e modelos treinados por modelos de aprendizado de máquina pré-treinados. Além do módulo de agregação de funções de rótulo. Entretanto, as funções desenvolvidas eram específicas para entidades na língua inglesa e quando aplicado a um *corpus* de Licitação e Contratação Pública em português, os resultados foram insatisfatórios.	Inspirada nessa proposta, esta pesquisa produziu uma arquitetura focada na construção de várias funções que atendesse a necessidade do contexto de Licitação e Contratação Pública.

Tabela 3.1: Continua na próxima página

(Luz de Araujo et al., 2018)	LeNER-Br	594 decisões (62.933 frases; 1.782.395 *tokens*; 33.055 anotações de granularidade mais ampla e 57.573 anotações de granularidade fina)	*"ORG"*, *"PER"*, *"LOC"*, *"TEMPO"*, *"LEGISLAÇÃO"*, *"JURISPRUDÊNCIA"*	Bi-LSTM-CRF	97.04% - Legislação e 88.82% - Entidades Legais;	A rotulação do corpus foi realizada manualmente por estudantes de direto compondo um corpus de entidades legislativas brasileiras	No trabalho de Luz de Araujo et al. (2018) não foi aplicado mecanismos para atribuição automatizada dos rótulos, dependendo na sua completude de esforço manual, diferentemente desta pesquisa, que não exige a rotulação manual.
(Albuquerque et al., 2022)	UlyssesNER-Br	TCU (371 votos, 44 jurisprudências, 4 áreas, 27 temas e 38 subtemas), STJ (7403 acórdãos, 1458 jurisprudências, 7 matérias e 68 naturezas)	*"PESSOA"*, *"DATA"*, *"EVENTO"*, *"ORGANIZAÇÃO"*, *"LOCAL"*, *"FUNDAMENTO'* e *"PRODUTODELEI"*.	CRF e Bi-LSTM	respectivamente, 80,8% na análise por categorias e 81,04% na análise por tipos; 76,89% na análise por categorias e 59,67% na análise por tipos	A rotulação do corpus foi realizada manualmente por especialista dos órgãos, compondo um corpus de entidades dos Acórdãos e Jurisprudências do TCU e STJ	No trabalho de Albuquerque et al. (2022) não foi aplicado mecanismos para atribuição automatizada dos rótulos, dependendo na sua completude de esforço manual, diferentemete desta pesquisa, que não exige a rotulação manual.

Capítulo 4

Metodologia adotada para rotulação por Supervisão Fraca

Embora a presença de um especialista no domínio seja extremamente vantajoso e poder resultar em melhores resultados para adotar a SF na rotulação de dados, ressalta-se que é possível entender os conceitos, padrões e nuances do domínio específico a partir do estudo do assunto e assim realizar a escolha das técnicas adequadas de SF, na seleção dos rótulos parciais mais relevantes e na interpretação dos resultados do modelo. Portanto, neste capítulo, daremos destaque à metodologia estratégica aplicada para a compreensão do domínio e para o desenvolvimento dos experimentos realizados neste trabalho.

Este trabalho propõe a criação de um corpus de Licitação e Contratação Pública anotado automaticamente a partir de métodos de SF (Lison et al., 2021). Esta pesquisa é composta pelo desenvolvimento de um conjunto de FR utilizando três mecanismos, correspondência por palavra, conhecimento heurístico (regras) e modelos pré-treinados por aprendizado de máquina, além da adaptação de um módulo de agregação (Seção 2.6.2).

A Figura 4.1 apresenta a metodologia utilizada na realização do trabalho. A primeira etapa da metodologia refere-se ao entendimento do problema a ser tratado, conforme elencando no capítulo 1 e na Seção 4.1. Na segunda etapa foi realizada uma revisão da literatura para compreensão da tarefa de NER e da abordagem de SF, conforme apresentado no Capítulo 2 e na Seção 4.2, por último foram realizados experimentos iniciais para melhor entendimento da tarefa de NER e do funcionamento dos *frameworks* que implementaram as diretrizes da SF. Na terceira etapa foi realizado o levantamento dos atos e das entidades que compunham o processo de contratação pública. Após este levantamento, foi feito a seleção de entidades que atendiam ao contexto da pesquisa, e por último, realizado o mapeamento e descrição das regras a serem implementadas. Na quarta etapa foi realizada a implementação das FR, a adaptação do módulo de agregação do Skweak(Lison et al., 2021) e o treinamento dos modelos de NER. A seguir cada uma das etapas são detalhadas.

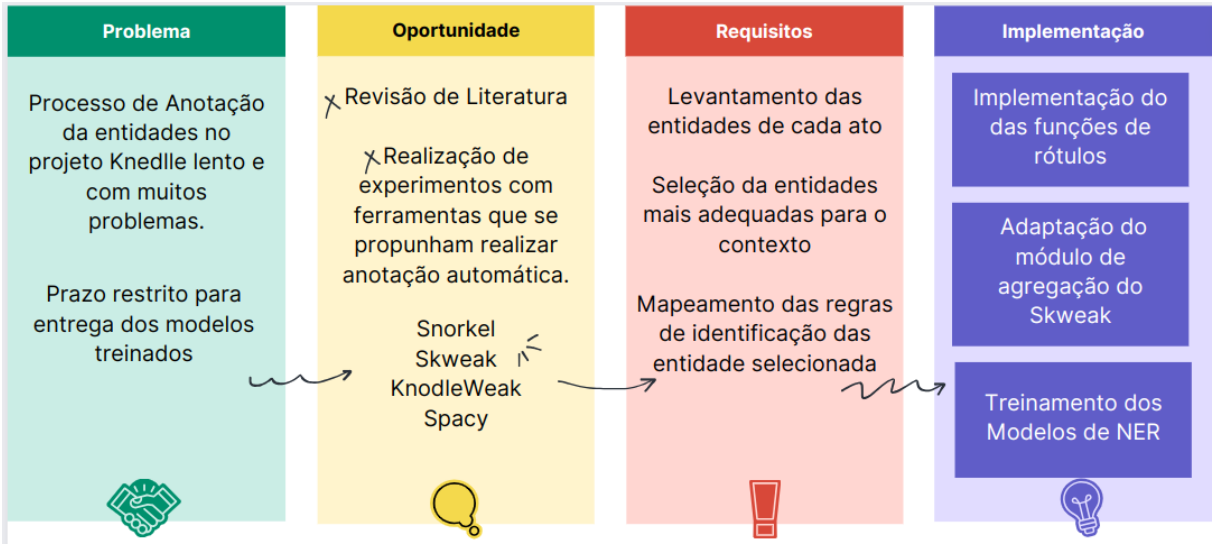


Figura 4.1: Visão Geral da Metodologia de Desenvolvimento do Trabalho.

4.1 Problema

Nas fases iniciais do projeto KnEDLe, houve um esforço considerável para a anotação dos atos. O conjunto de dados utilizado para anotação foi composto por fragmentos de dados extraídos do DODF na *Seção II* e *Seção III* de julho de 2021 a julho de 2022. Os dados do projeto KnEDLe foram coletados em formato PDF e convertidos em TXT por meio da ferramenta *DODFMiner*. *DODFMiner* é uma ferramenta de extração de dados projetada especificamente para extrair dados estruturados do DODF. Dessa forma, após a extração dos dados, um conjunto de cem documentos foram definidos para anotação manual. O processo de anotação foi executado por meio de uma combinação de anotadores voluntários e pesquisadores que receberam treinamento de especialistas no domínio do conhecimento. Após o processo de anotação, cada documento anotado foi revisado por outro anotador treinado por um especialista.

A Figura 4.2 mostra um exemplo de anotação feita pelos anotadores. Todo o parágrafo é destacado como entidade do tipo de ato publicado, enquanto cada segmento destacado com uma cor diferente é uma entidade que pertence ao ato.

Durante o processo de anotação, um grupo de "20%" dos documentos do DODF, denominado "lote de validação", foi usado para avaliar a confiabilidade dos rótulos e a qualidade dos textos anotados utilizando as métricas de concordância do conjunto de dados. A versão final do conjunto de dados KnEDLe inclui informações sobre atos de pessoal e atos de licitação pública e foi validado por membros do projeto KnEDLe, podendo ser utilizado como um conjunto de dados CPO. O CPO está disponível no repositório do projeto [1]. Vale ressaltar que esses são dados públicos e abertos e podem ser acessados por qualquer pessoa.

A fim de demonstrar a onerosidade do processo de anotação manual, e reforçar a importância de encontrar alternativas para substituir esse processo, pode-se citar uma série de fatores que afetaram o processo de anotação, tais como:

[1]Disponível em `https://github.com/UnB-KnEDLe/datasets/blob/master/corpus_2_contratos_licitacoes.md`

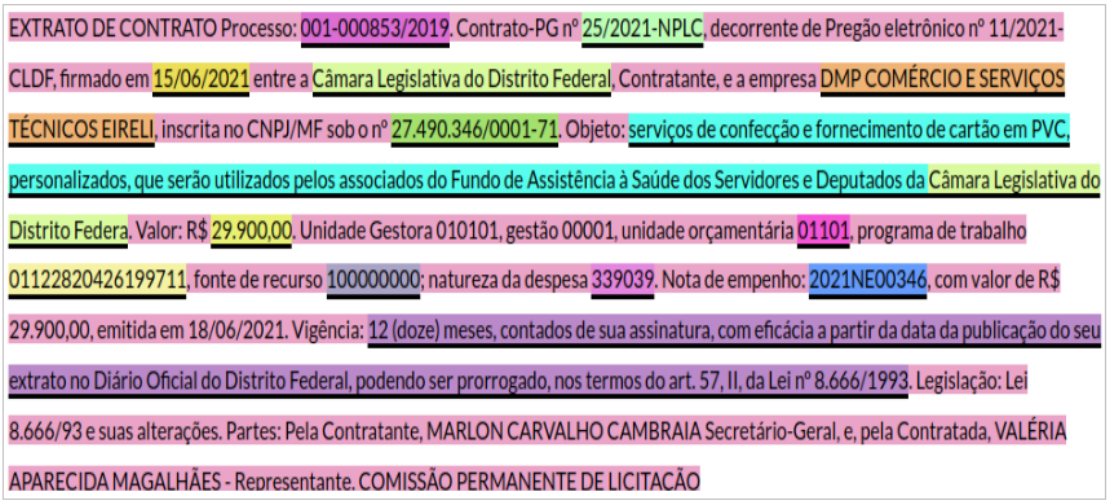


Figura 4.2: Exemplo de uma anotação com as entidades destacadas Schmidt (2019)

- falta de anotadores disponíveis para realizar a tarefa de anotação;
- falta de especialização entre os anotadores voluntários no idioma jurídico;
- falta de familiaridade com as entidades e atos apresentados no DODF;
- interrupções causadas por atividades de pesquisa e questões de saúde entre alguns anotadores;
- necessidade de priorizar outras atividades do projeto.

Portanto, o processo de rotulação manual se deu de forma lenta e onerosa, levando ao todo em torno de um ano e meio para conclusão.

Esses desafios colocaram em risco os prazos estabelecidos para o projeto, já que não seria possível avançar para as próximas etapas sem um *corpus* anotado. A fim de evitar essa situação em projetos futuros, os gestores propuseram várias alternativas de abordagens semi-supervisionadas e não supervisionadas. Entre essas alternativas, o aprendizado fracamente supervisionado por meio da adoção das técnicas de SF foi uma das selecionadas para ser adotada como instrumento de anotação automática, utilizando como escopo os atos de Licitação e Contratação Pública.

A oportunidade de investigar a aplicação da abordagem SF justificou-se pelo fato dos atos da Seção III, Licitação e Contratação Pública, possuírem uma grande variedade de entidades em comum. Por esse motivo, o uso da SF é ideal para treinar modelos de NER com rótulos mais genéricos ou abrangentes, como identificar palavras-chave ou padrões associados às entidades, em vez de rótulos precisos para cada ato. Outra vantagem da utilização dessa abordagem nestes tipos de entidades é a reutilização de modelos pré-treinados. Em vez de treinar modelos de NER do zero, a SF pode ser usada para pré-treinar modelos em um conjunto específico de atos da Seção III, os quais podem então serem ajustados ou combinados para outros atos da mesma seção.

4.2 Oportunidades

A proposta do modelo iniciou-se após realização da revisão da literatura conforme descrito no Capítulo 3 e execução de experimento preliminares para entendimento da tarefa de NER e para entendimento da aplicação prática das técnicas de SF. Como exemplos dessas técnicas pode-se citar: O Snorkel (Bach et al., 2019), Skweak(Lison et al., 2021), KnodleWeak (Sedova et al., 2021) e o Spacy (Honnibal and Montani, 2017).

No início desse estudo a anotação da base de dados de licitação e contratos ainda estava em andamento. A fim de avançar com a pesquisa, optou-se por utilizar bases conhecidas na literatura tais como *CoNLL* (Sang and De Meulder, 2003) e Ontontes 5 (Weischedel, 2013) para realização dos experimentos iniciais. Esses experimentos iniciais serviram de fundamento para as decisões de pesquisas e experimentações descritas nas próximas seções.

4.3 Requisitos

Durante a fase de requisitos, foi conduzido um levantamento das entidades associadas a cada ato. Após essa etapa, realizou-se um estudo abrangente de todas as entidades, seguido pela seleção das que seriam utilizadas nos experimentos. Essa seleção foi feita visando explorar diversos cenários na implementação das FR, como detalhado na Seção 4.3.2. Em seguida, procedeu-se com o mapeamento e documentação dos padrões e características de cada entidade escolhida.

4.3.1 Levantamento das Entidades

Essa etapa teve como objetivo realizar o levantamento das entidades associadas a cada ato praticado pelo poder público no processo de licitação e contração pública. Inicialmente, foi realizado um estudo de como funcionava o processo de contratação pública e quais atos eram praticados pelo TCDF. O processo de licitação e contratação pública é um processo público para a realização de compras de produtos e contratação de serviços e de obras públicas, juridicamente norteadas pela Lei de Licitações e Contratos Administrativos, sob o número 14.333/2021[2]. Os atos praticados nesse processo são:

- Aviso de Suspensão de Licitação
- Aviso de Revogação/Anulação de Licitação
- Extrato de Contrato ou Convênio
- Extrato de aditamento contratual
- Aviso de Abertura de Licitação

A figura 2.2 demostra um exemplo do ato *"Extrato de Contrato"* e um exemplo de uma entidade associada a este ato. Neste caso se destacou a entidade *"Data de Assinatura"* :

O levantamento completo dos atos e suas respectivas entidades estão presentes no Apêndice A.

[2]Disponível em `http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/L14133.htm`

97,"EXTRATO DO CONTRATO No 37375/SEDICT/DF PROCESSO SEI-GDF no 00370-00001322/2018-13. DAS PARTES: SECRETARIA DE ESTADO DE ECONOMIA, DESENVOLVIMENTO, INOVACAO, CIENCIA E TECNOLOGIA (CONTRATANTE) e NEANTRO SAAVEDRA RIVANO, CPF/MF no 592.374.577-15 (CONTRATADO). OBJETO: Prestacao de servicos de consultoria individual. DA VIGENCIA: 120 (cento e vinte reais) dias a contar da assinatura. DO VALOR: R$ 74.400,00 (Setenta e quatro mil e quatrocentos reais). MODALIDADE: Ordinario. DA DOTACAO ORCAMENTARIA: Elemento de Despesa: 3.3.90.39 - Projeto/Atividade/Programa de Trabalho: 22661620750210001 - Modernizacao e Melhoria da infraestrutura das Areas de Desenvolvimento Economico do DF. DA NATUREZA DA DESPESA: 33.90.35 - servicos de consultoria. DA FONTE DE RECURSO: 136008662. DA NOTA DE EMPENHO: 2018NE00345. DA UNIDADE ORCAMENTARIA: 20101. DO PROCEDIMENTO: Licitacao Publica, com base na Lei 8.666/93. DA LEGISLACAO: Contrato de Emprestimo No 2957/OC-BR celebrado entre o Governo e o BID; Lei Federal no 8.666, de 21/06/93, com as alteracoes subsequentes e demais legislacoes pertinentes a materia. DATA DE ASSINATURA: 02/10/2018. SIGNATARIOS: Pelo Contratante, ANTONIO VALDIR OLIVEIRA FILHO, na qualidade de Secretario de Estado e, pelo Contratado, NEANTRO SAAVEDRA RIVANO na qualidade de consultor individual."

Figura 4.3: Texto de Extrato de Contrato

4.3.2 Seleção das Entidades

A partir do levantamento das entidades, foi possível analisar e selecionar as entidades que seriam utilizadas como experimentos desses estudo, e dessa forma realizar a implementação das FR. Para selecionar as entidades utilizou-se:

- Quais entidades estavam presentes em mais de um tipo de ato?
- Quais entidades anotadas com NER tradicional estavam presentes em mais de um tipo de ato e possuíam resultados de alta precisão e alta revocação no modelo treinado?
- Quais entidades anotadas com NER tradicional apresentaram resultados de baixa precisão e/ou baixa revocação?
- Quais entidades apresentavam padrões de publicação bem definidos que eram fortes candidatas para mapeamento heurístico e definição de regras?
- Quais entidades apresentavam um conjunto fechado de valores permitindo criar uma lista desses rótulos para realizar a correspondência por palavras?

Como resposta a essas questões obteve-se a seguinte lista de entidades por ato:

- **Extrato de Contrato**: data da assinatura do contrato, tipo de despesa, nota de compromisso, número do contrato, processo GDF, programa de trabalho, unidade orçamentária e valor do contrato
- **Extrato de Convênio**: data da assinatura do convênio, processo GDF e número do convênio.
- **Suspensão de Contrat**o: processo GDF, número da licitação e modalidade de licitação.
- **Anulação e Revocação da Licitação**: processo GDF, número da licitação, modalidade de licitação e data escrito.
- **Aviso de Licitação**: processo GDF, número da licitação, modalidade de licitação e valor estimado para contratação.
- **Aditamento Contratual**: data da assinatura do contrato, processo GDF e número do contrato.

4.3.3 Mapeamento das Entidades

Após a seleção das entidades, procedeu-se com o mapeamento e a documentação das características de cada uma delas, com base nos trechos publicados por tipo de ato.

Para cada entidade, foram catalogados pelo menos 10 exemplos de diferentes tipos de ocorrências nos textos dos diversos atos. Um exemplo dessa documentação pode ser visualizado na Figura 4.4:

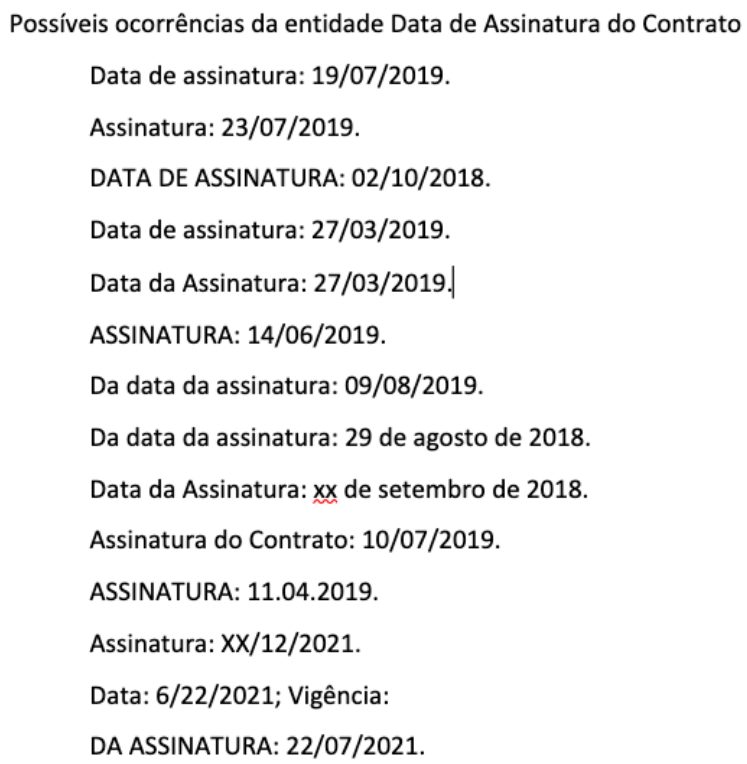


Figura 4.4: Exemplo de ocorrência da entidade *"Data de Assinatura" no ato Extrato de Contrato*

4.3.4 Definição das funções de rótulo

Após a seleção das entidades, foi feito um estudo de quais tipos de FR poderiam ser implementadas de acordo com cada particularidade das entidades mapeadas. Por exemplo, entidades que apresentavam uma estrutura de formação passível de ser identificada por meio de regras do tipo expressão regular. Neste caso, foi realizado um mapeamento dos padrões adotados no processo de anotação manual, os quais foram descritos em linguagem natural, de forma que fosse possível codificar esse padrão. Dessa forma, as entidades *"data da assinatura do contrato"*, *"tipo de despesa"*, *"nota de compromisso"*, *"número do contrato"*, *"processo GDF"*, *"programa de trabalho"*, *"unidade orçamentária"*, *"valor do contrato"*, *"data da assinatura do acordo"*, *"número do acordo"*, *"número da licitação"*, *"modalidade da licitação"*,*"data escrita"*, *"valor estimado para contratação"* foram selecionadas para serem utilizadas na implementação de uma função de rótulo do tipo conhecimento heurístico (regras).

Da mesma forma, foram identificadas algumas entidades que apresentavam uma variedade de domínios pré-definidos, tais como: *"unidade orçamentária"* e *"modalidade do processo de licitação"*. Sendo fortes candidatas para serem utilizadas nas FR do tipo correspondência por palavra.

Algumas entidades foram treinadas em outros modelos, alcançando boas precisões e revocações. Assim, algumas dessas entidades, tais como *"números da licitação"* e *"processos GDF"* foram selecionadas para serem utilizadas na aplicação da função de rótulo de modelos de aprendizado de máquina pré-treinados para extrair rótulos das entidades que a acurácia da rotulação alcançada pelas funções de conhecimento heurístico (regras) e correspondência por palavras foram baixas.

4.4 Implementação

Nesta seção, são descritos os detalhes da implementação realizada para aplicar a abordagem da SF no contexto desta pesquisa. A arquitetura da implementação foi divida em duas etapas: etapa de implementação das FR, adaptação do modelo de agregação e etapa de treinamento dos modelos.

Para a implementação das FR baseadas na abordagem de SF foi utilizada a linguagem Python e parte do *framework* Skweak (Lison et al., 2021), que além de facilitar o desenvolvimento, execução e validação das FR, disponibiliza o módulo agregação. Para tokenização do texto, adotou-se o *Tokenizer* do *Spacy* [3]

Um dos fatores que contribuíram na escolha do Skweak como *framework* foi o fato dele ser escrito em Python e oferecer o módulo de agregação, uma vez que o restante da aplicação também foi codificada usando esta linguagem. Python é uma linguagem simples e de fácil utilização, permite desenvolvimento orientado a objetos e possui uma vasta gama de bibliotecas de códigos, principalmente na área estatística. Para esta pesquisa utilizou-se, dentre outras, as bibliotecas numpy, pandas e scikit-learn.

A escolha dos modelos de treinamento seguiu alguns critérios estabelecidos, sendo selecionados dois modelos: *Sklearn-CRF* e o Bi-LSTM com uma camada CNN.

4.4.1 Implementação das funções de rótulos e Adaptação do Módulo de Agregação

Foi implementado um conjunto extensivo de FR para alcançar uma boa precisão da anotação da base de supervisão fraca. Conforme citado na Seção 2.6.1, esta pesquisa focou em basicamente três tipos de FR: correspondência por palavra, conhecimento heurístico (regras) e modelos de aprendizado de máquina pré-treinados.

O código *python* 4.1 demonstra um exemplo de função de rótulo do tipo regra, implementado para identificar a entidade *"Data de Assinatura"* do ato *"Extrato de Contrato"* :

[3]Disponível em `https://spacy.io/api/tokenizer`

[3]Disponível em `https://sklearn-crfsuite.readthedocs.io/en/latest/api.html`

[3]Disponível em `https://pytorch.org/tutorials/beginner/nlp/advanced_tutorial.html`

```
class LabelFunctionsContrato:
    def __init__(self, dados):
        nlp = spacy.load('pt_core_news_sm', disable=["ner","lemmatizer"])
        self.docs = list(nlp.pipe(dados))
    def data_assinatura_(self, doc):
        '''
        label function para extracao de data de assinatura usando regras
        parametros:
            doc: uma string respresentando o texto de um dos contratos
            oferecidos no vetor da base de dados
        '''

        expression =
        r[A|a][S|s][S|s][I|i][N|n][A|a][T|t][U|u][R|r][A|a].*?[\s\S]
        (\d{2}\/\d{2}\/\d{4}|\d{2}[\s\S]\w+[\s\S]\w+[\s\S]\w+
        [\s\S]\d{4}|\w{2}[\s\S]\w+[\s\S]\w+[\s\S]\w+[\s\S]\d{4}|[\s\S]
        (\d{2}\.\d{2}\.\d{4})|\w{2}/\d{2}\/\d{4})

        match = re.finditer(expression, str(doc))
        if match:
            for grupo in match:
                flag = 0
                for token in doc:
                    if grupo.span(1)[0]+1 in range(token.idx,
                    (token.idx+len(token))+1) and flag == 0:
                        if(doc[token.i].text == ':'):
                            start = doc[token.i+1]
                        else:
                            start = token
                        flag = 1
                    if token.idx >= grupo.span(1)[1] and flag == 1 and
                    token.i > start.i:
                        %if(doc[token.i-1].text in ['.', '-', '-', ':']):
                            end = doc[token.i-1]
                        else:
                            end = token
                        yield start.i, end.i, "data_assinatura_contrato"
                        break
```

Source Code 4.1: Função de rótulo de regra para entidade *"Data de Assinatura"* do ato *"Extrato de Contrato"*

O código *python* 4.2 demonstra um exemplo de função de rótulo do tipo correspondência por palavra, implementado para identificar a entidade *"Nota de Empenho"* do ato *"Extrato de Contrato"*:

```
class LabelFunctionsContrato:
    def __init__(self, dados):
        nlp = spacy.load('pt_core_news_sm', disable=["ner","lemmatizer"])
        self.docs = list(nlp.pipe(dados))
   def nota_emp_detector_fun(self, doc):
        '''
        label function para extracao de nota de empenho com comparacoes
        de listas

        parametros:
            doc: uma string respresentando o texto de um dos contratos
            oferecidos no vetor da base de dados
        '''
        for token in doc:
            if token.i+2 < len(doc):
                for y in ['Empenho', 'EMPENHO', 'Empenho:', 'EMPENHO:',
                'empenho:']:
                    if y in token.text:
                        k = 0
                        if((len(doc[token.i+1].text) <= 2 or
                        doc[token.i+1].text in
                        ['No', 'NO', 'no', 'Nº', 'nº', 'N°', 'n°', 'n',
                        'n.', 'n.º'])):
                            if 'R$' in doc[token.i+1].text:
                                break
                            k += 1
                        if(k >= 1 and (len(doc[token.i+2].text) <= 2 or
                        doc[token.i+2].text in
                        'No', 'NO', 'no', 'Nº', 'nº', 'N°', 'n°', 'n',
                        'n.', 'n.º'])):
                            if 'R$' in doc[token.i+2].text:
                                break
                            k += 1
                        if(doc[token.i+1+k].text.isalpha()):
                            break
                        for x in range(1+k, len(doc)-token.i):
                            if 'R$' in doc[token.i+x].text:
                                break
                            if (doc[token.i+x].text.isalpha() or
                            doc[token.i+x].text in ['.', ',', ';'])
                            and token.i+1+k < token.i+x:
                                yield token.i+1+k, token.i+x,
                                "nota_empenho"
                                break
                            elif token.i+x+1 >= len(doc) and token.i+1+k<
                            token.i+x+1:
```

```
                                yield token.i+1+k, token.i+x+1,
                                "nota_empenho"
                                break
```

Source Code 4.2: Função de rótulo de correspondência para entidade *"Nota de Empenho"* do ato *"Extrato de Contrato"*

O código *python* 4.3 demonstra um exemplo de função de rótulo do tipo modelos de aprendizado de máquina pré-treinados, implementado para identificar a entidade *"Processo GDF"* do ato *"Extrato de Convênio"* :

```
class LabelFunctionsConvenio:
    def __init__(self, dados):
        nlp = spacy.load('pt_core_news_sm', disable=["ner","lemmatizer"])
        self.docs = list(nlp.pipe(dados))
    def processo_ml_fun(self, doc):
            #Extrai as features do CRF
            def _get_features(sentence):
                """Create features for each word in act.
                Create a list of dict of words features
                to be used in the predictor module.
                Args:
                    act (list): List of words in an act.
                Returns:
                    A list with a dictionary of features
                    for each of the words.
                """
                sent_features = []

                for i in range(len(sentence)):
                    word_feat = {
                        'word': sentence[i].lower(),
                        'capital_letter': sentence[i][0].isupper(),
                        'all_capital': sentence[i].isupper(),
                        'isdigit': sentence[i].isdigit(),
                        'word_before': sentence[i].lower()
                        if i == 0 else sentence[i-1].lower(),
                        'word_after:': sentence[i].lower()
                        if i+1 >=len(sentence) else sentence[i+1].lower(),
                        'BOS': i == 0,
                        'EOS': i == len(sentence)-1
                    }
                    sent_features.append(word_feat)
                return sent_features
```

```
            # Carregar modelo
            with open('./crf_modelo_extrato_contrato.pkl', 'rb') as f:
                model = pickle.load(f)
            validacao = []

            for token in doc:
                validacao = _get_features(token.text)

            processo_lb = model.predict(validacao)

            resultados = []
            comeco_tag = 0
            final_tag = 0

            #pega a posicao do token de inicio e fim
            for token_list, tag_list in
            zip(token.text, processo_lb):
                for (idx_token, token),(idx_tag, tag)
                in zip(enumerate(token_list), enumerate(tag_list)):
                    if tag == 'B-processo_gdf':
                        acumulador = 0
                        comeco_tag = idx_tag
                        for tag_seguinte in tag_list[idx_tag+1:]:
                            if tag_seguinte == 'I-processo_gdf':
                                acumulador+=1
                            else:
                                break
                        final_tag = comeco_tag + acumulador
                        yield comeco_tag, final_tag, 'processo_gdf'
                        break
```

Source Code 4.3: Função de rótulo modelos de aprendizado de máquina pré-treinados para a entidade *"Proceso GDF"* do ato *"Extrato de Convênio"*

Após aplicar as FR ao *corpora* é necessário agregar os resultados para obter uma única anotação probabilística (em vez das múltiplas anotações possivelmente conflituosas das FR). Esse processo é feito com adoção do módulo de agregação do framework Skweak (Lison et al., 2021), o qual é baseado em um modelo generativo, que estima automaticamente a precisão relativa e possíveis ruídos de cada função, escolhendo de forma estatística o rótulo a ser atribuído a entidade.

O código *python* 4.4 demonstra o uso do módulo de agregação do Skweak adaptado para aplicação no contexto desta pesquisa.

```
class LabelFunctionsContrato:
    def __init__(self, dados):
        nlp = spacy.load('pt_core_news_sm', disable=["ner","lemmatizer"])
        self.docs = list(nlp.pipe(dados))
    def train_HMM_Dodf(self):
        '''
        treina o modelo HMM para refinar e agregar a entidades extraidas
        pelas label functions
        '''

        model = skweak.aggregation.HMM("hmm",
        ["numero_contrato", "processo_gdf", "valor_contrato",
        "unidade_orcamentaria", "programa_trabalho",                    "n
        "data_assinatura_contrato"], sequence_labelling=True)

        self.docs = model.fit_and_aggregate(self.docs)

        for doc in self.docs:
            if "hmm" in doc.spans:
                doc.ents = doc.spans["hmm"]
            else:
                doc.ents = []

        ''' Salvando modelo HMM em uma pasta data '''
        if os.path.isdir("./data"):
            skweak.utils.docbin_writer(self.docs,
            "./data/reuters_small.spacy")
        else:
            os.mkdir("./data")
            skweak.utils.docbin_writer(self.docs,
            "./data/reuters_small.spacy")
```

Source Code 4.4: Módulo de Agregação

A lista completa das FR implementadas, bem como os demais códigos utilizados na realização desta pesquisa estão disponíveis no repositório do projeto [4] .

4.4.2 Modelos para o Reconhecimento de Entidade Nomeada

Nesta seção são apresentados os modelos utilizados para realização do treinamento dos dados. A fim de alcançar o objetivo proposto, foram selecionados alguns modelos capazes de realizar a tarefa de NER a partir da base gerada com a SF em vários idiomas, com

[4]Disponível em `https://github.com/UnB-KnEDLe/experiments/tree/master/members/lucelia/Supervisao_Fraca`

foco particular no idioma português. Os principais critérios para a escolha dos modelos foram:

- Soluções totalmente gratuitas
- Soluções independentes de idioma e domínio
- Soluções que permitissem treinar um modelo customizado para NER, com classes de entidades customizadas.

Os modelos de NER escolhidos foram utilizados em dois momentos: para treinamento do modelo a partir da criação da base de supervisão fraca e para treinamento dos modelos a partir da base ouro.

Conforme os critérios estabelecidos, foram selecionados dois modelos: *Sklearn-CRF* e o Bi-LSTM com uma camada CNN.

O *Sklearn-CRF* apesar de um algoritmo antigo, é considerado um dos melhores para os problemas de NER. Com o CRF é possível alcançar alta qualidade de rotulação, ao implementar a definição de características a serem utilizadas na identificação do rótulo. Também é possível ter flexibilidade o suficiente em termos de seleção de recursos. Além disso, em relação a desempenho, o CRF, por não ser baseado em DL, é mais rápido e não exige muito recurso computacional, sendo possível realizar os experimentos em máquinas convencionais.

O Bi-LSTM, por utilizar duas de redes neurais independentes permite que a execução da entrada de dados seja realizada tanto de frente para trás ou de trás para frente, fornecendo assim melhores resultados da predição, uma vez que possui um melhor entendimento do contexto, mostrando assim ser melhor para problema de NER em detrimento a outros algoritmos. O Bi-LSTM, apesar de exigir mais recursos computacionais que o *Sklearn-CRF*, mostrou-se que é um modelo barato em relação à memória e tempo de processamento.

Além dos critérios acima, os modelos foram escolhidos por serem vastamente utilizados pela academia e são modelos de referência utilizados nos trabalhos de Lison et al. (2021) e Ratner et al. (2017), os quais possuem um foco especial em trabalhar com *frameworks* que não exigem quaisquer dados anotados à mão. Em vez de anotar dados manualmente, utilizaram-se as diretrizes da SF para rotular programaticamente os dados mediante uma coleção de FR.

[4]Disponível em `https://sklearn-crfsuite.readthedocs.io/en/latest/api.html`

[4]Disponível em `https://pytorch.org/tutorials/beginner/nlp/advanced_tutorial.html`

Capítulo 5

Avaliação De Desempenho

Com o intuito de validar a proposta metodológica, descrita na Seção 1.2 foi realizada a avaliação de desempenho de dois experimentos. No Experimento 1, Seção 5.2.1 foi realizado o treinamento de modelos NER com o CPO e com a base de dados gerada pela abordagem da SF. Já no Experimento 2, Seção 5.2.2, foi avaliado a aplicação das FR diretamente ao conjunto de dados e sem realizar o treinamento de modelos para NER.

A Seção 5.1 apresenta a base de dados utilizados nos experimentos. A seleção dos modelos de NER está descrita na Seção 4.4.2. A metodologia utilizada na avaliação é apresentada na Seção 5.3. Já a execução do Experimento 1 e Experimento 2 estão descritos na Seção 5.2.

5.1 Base de dados

Esta pesquisa visa investigar se as técnicas de SF podem facilitar a geração rápida de dados anotados e a criação de modelos com qualidade comparável aos modelos criados por humanos, resultando em uma redução nos custos de rotulação de dados.

Portanto, para realizar a validação dos resultados obtidos com o treinamento sobre a base gerada pela SF foi utilizado um conjunto de dados em português e de domínio legal específico para o problema de NER. Esse conjunto de dados foi anotado pela equipe do projeto KnEDLe, conforme pode ser melhor compreendido na Seção 4.1. Este conjunto de dados contempla os seis tipos atos celebrados entre Tribunal de Contas do Distrito Federal e as instituições contratadas, e estão disponíveis no repositório do projeto [1] para acesso público.

A base de dados de atos de contratos possui um total de 4090 atos anotados. A Tabela 5.1 descreve os tipos de atos considerados, o número de atos anotados, a quantidade de entidades e a quantidade de palavras por tipo de ato.

5.1.1 Pre-Processamento dos Dados

A primeira versão da base de dados de contrato e licitações liberada para uso apresentava duplicidade de chaves de alguns atos, as quais provocavam erro na aplicação do treinamento de NER. Dessa forma, foi necessário reunir com o pessoal da anotação para que

[1] Disponível em `https://github.com/UnB-KnEDLe/datasets`

Tabela 5.1: Base Ouro de Licitações e Contrato V2

Atos	Qtde de atos	Qtde de Entidades	Qtde de palavras
EXTRATO CONTRATO	1734	21432	324718
ADITAMENTO CONTRATO	1551	10293	278730
AVISO LICITACAO	639	6790	88860
SUSPENSAO LICITACAO	82	503	9057
ANUL REVOG LICITACAO	52	345	5573
EXTRATO CONVENIO	32	280	5715
Total	4090	43.733	712.653

fossem eliminadas estas duplicidades. A base também apresentavam quebras de sentença, espaços em brancos e caracteres especiais indevidos, sendo necessário realizar uma limpeza dos dados.

Após a limpeza, foi aplicada a marcação de estilo *IOB* padrão palavra por palavra que captura as mesmas informações para uma entidade. Para realizar a *tokenização* dos dados foi utilizado o *Tokenizer* do *Spacy* [2] a fim de padronizar a *tokenização* do conjunto de treinamento e validação utilizado para avaliação dos modelos treinados com SF.

As informações descritivas sobre a base dados, tais como as entidades e a quantidade de cada entidade por tipo ato, podem ser obtidas nas tabelas B.1, B.2, B.3, B.4, B.5, B.6 presentes no Apêndice B

5.2 Execução dos Experimentos

A questão de pesquisa deste estudo indaga como gerar modelos automatizados e precisos para a tarefa de NER, especialmente quando o contexto em questão possui poucos dados anotados disponíveis em português e os conjuntos de dados existentes exibem características contextuais específicas que requerem o envolvimento de pessoal especializado na anotação de dados. Em resposta a essa questão, foi examinada a hipótese de que a técnica de SF pode contribuir para a rápida geração de dados anotados e para a criação de modelos com qualidade comparável aos modelos desenvolvidos com dados anotados por humanos. Para investigar essa hipótese, foram realizados experimentos que envolvem o treinamento de NER tanto no CPO quanto na base gerada pela SF. Além disso, foi conduzida uma análise comparativa de desempenho entre os rótulos gerados exclusivamente pela aplicação da SF, sem a necessidade de treinamento de NER.

Todos os experimentos foram segmentados por tipo de ato de licitação e contrato, os quais estão descritos na Seção A. O treinamento foi realizado para cada ato em ambas as bases. Da mesma forma, uma comparação do desempenho de rótulos gerados exclusivamente com a aplicação de SF foi realizada para cada ato.

[2]Disponível em `https://spacy.io/api/tokenizer`

5.2.1 Experimento 1 - Treinamento dos Modelos de Reconhecimento de Entidade Nomeada

Este experimento foi conduzido para avaliar o desempenho de modelos NER treinados sobre o CPO e sobre base gerada pela SF. O objetivo desse experimento foi avaliar se os resultados alcançados pelo modelo de NER sobre a base de SF eram superiores aos resultados alcançados pelos modelos de NER tradicional sobre o CPO.

Os modelos usados para o treinamento de NER são os mencionados na Seção 4.4.2.

Experimento 1.1 - Treinamento de NER usando o modelo *Sklearn-CRF*

O Experimento 1.1 aborda o treinamento de NER utilizando o modelo *Sklearn-CRF* sobre o CPO e sobre a base SF.

Para realizar esses treinamentos, foram definidas diversas características, tais como, se a palavra era escrita em letras minúsculas ou maiúsculas, se era um título, se continha dígitos, o sufixo, a *postag* e outras informações próximas à palavra. A Figura 5.1 mostra o vetor de recursos criados para cada *token* do texto de entrada definido para o treinamento do CRF.

```
[{'bias': 1.0,
  'word.lower()': 'rejects',
  'word[-3:]': 'cts',
  'word[-2:]': 'ts',
  'word.isupper()': False,
  'word.istitle()': False,
  'word.isdigit()': False,
  'postag': 'VBZ',
  'postag[:2]': 'VB',
  'BOS': True,
  '+1:word.lower()': 'german',
  '+1:word.istitle()': True,
  '+1:word.isupper()': False,
  '+1:postag': 'JJ',
  '+1:postag[:2]': 'JJ'},
```

Figura 5.1: Exemplo de saída após extração das características

Da mesma forma, para ambos treinamentos, foi utilizado o algoritmo de descida do gradiente L-BFGS (*Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm*), com número máximo de iteração igual a 70 e os coeficientes de regularização *Eslastic Net* (L1 e L2). Com intuito de otimizar os parâmetros utilizados nos coeficientes de regularização foi realizada um cruzamento sobre as configurações por meio do *RandomizedSearchCV* do *sklearn* com os seguintes parâmetros: CRF, *params_space*, *cv=5*, *verbose=1*, *n_jobs=-1*, *n_iter=70*, *scoring=f1_scorer*.

Para cada ato foram utilizados os seguintes coeficientes de regularização: (1) extrato do contrato (C1=0,1094, C2=0,0230), (2) extrato do convênio (C1=0,1094, C2=0,0230), (3) aviso de licitação (C1=0,0629, C2=0,0510), (4) suspensão da licitação (C1=0,461, C2=0,05), (5) anulação/revogação da licitação (C1=0,1181, C2=0,0069) e (6) Aditamento Contratual (C1=0,0773, C2=0,1447).

A seguir são apresentadas as etapas para expressar de forma macro as atividades realizadas em cada treinamento.

- Etapas realizadas para conduzir o treinamento NER sobre o CPO .

    1. Etapa: O processamento de dados foi realizado conforme descrito na Seção 5.1.1.
    2. Etapa: O modelo *Sklearn-CRF* foi treinado sobre o CPO, com aplicação da validação cruzada *K-fold* com cinco *folds*.
    3. Etapa: O modelo treinado foi aplicado às partições de teste para gerar rótulos preditos.
    4. Etapa: Os rótulos preditos gerados foram avaliados em relação aos rótulos originais do CPO.
    5. Etapa: Os resultados foram comparados conforme descrito na Seção 5.3.

- Etapas realizadas para conduzir o treinamento NER sobre a base de SF .

    1. Etapa: O processamento de dados foi realizado conforme descrito na Seção 5.1.1.
    2. Etapa: Os rótulos reais do CPO foram limpos.
    3. Etapa: A implementação das FR, descrita na Seção 4.4, foi aplicada ao CPO limpo para criar um banco de dados com rótulos fracos.
    4. Etapa: O modelo *Sklearn-CRF* foi treinado sobre a base de supervisão fraca gerada, com aplicação da validação cruzada *K-fold* com cinco *folds*.
    5. Etapa: As partições de teste utilizadas na validação do treinamento sobre a CPO foram recuperadas.
    6. Etapa: Os rótulos reais dessas partições de teste foram limpos.
    7. Etapa: O modelo treinado foi aplicado às partições de teste para gerar rótulos preditos.
    8. Etapa: Os rótulos preditos gerados foram avaliados em relação aos rótulos originais do CPO.
    9. Etapa: Os resultados foram comparados conforme descrito na Seção 5.3.

Experimento 1.2 - Treinamento de NER usando o modelo CNN Bi-LSTM

O Experimento 1.2 aborda o treinamento de NER utilizando o modelo CNN Bi-LSTM sobre o CPO e sobre base SF.

Para realizar esses treinamentos foram separadas as palavras e as classes, sobre as quais foi criado um dicionário contendo os índices de cada palavra. Na sequência, foi

substituído as palavras pelo seu índice no dicionário criado, sendo adicionado *tokens* de início e final de sentença para criar os *mini-batches* de treinamento.

Os melhores hiperparâmetros encontrados após o processo de otimização dos modelos, para cada tipo de ato, estão descritos na Tabela 5.2

Atos	*batch size*	*epochs*	*units*	*learning rate*
Extrato de Contrato	12	50	340	0.0030
Aviso de Licitação	5	50	140	0.0055
Aviso de Suspensão	5	50	200	0.0065
Aditamento Contratual	12	50	180	0.0010
Anulação e Revogação	5	50	105	0.0080
Extrato de Convenio	12	50	340	0.0030

Tabela 5.2: Hyperparâmetros Bi-LSTM

Foi utilizado ainda o algoritmo de otimização *Adam* [3] com a taxa de perda padrão de 0,0001, além da implementação da parada antecipada que será acionada se a perda de validação não diminuir em 5 épocas consecutivas.

A seguir são apresentadas as etapas para expressar de forma macro as atividades realizadas em cada treinamento.

- Etapas realizadas para conduzir o treinamento NER sobre o CPO .

    1. Etapa: O processamento de dados foi realizado conforme descrito na Seção 5.1.1.
    2. Etapa: O modelo CNN Bi-LSTM foi treinado sobre o CPO com aplicação da validação cruzada *K-fold* com cinco *folds* e com as seguintes camadas:
        - Camada de *embeddings* para transformar índices das palavras em vetores numéricos para que a rede neural aprenda representações semânticas mais significativas para as palavras
        - Camada *Convolutional Neural Network* Convi1D para extração de características e otimização da rede neural
        - Camada de *batchNormalization* para melhorar a estabilidade e a velocidade de treinamento da rede neural
        - Camada Bi-LSTM que permite que a rede neural processe as sequências de dados em duas direções simultaneamente
        - Camada de *dropout* para evitar o *overfitting*
        - Camada de *TimeDistributed* usada para aplicar uma camada densa em cada passo de tempo da sequência de entrada antes de alimentar os resultados para a camada LSTM
    3. Etapa: O modelo treinado foram aplicados às partições de teste para gerar rótulos preditos.
    4. Etapa: Os rótulos preditos gerados foram avaliados em relação aos rótulos originais do CPO.

[3]Disponível em `https://keras.io/api/optimizers/adam/`

5. Etapa: Os resultados foram comparados conforme descrito na Seção 5.3.

- Etapas realizadas para conduzir o treinamento NER sobre a base de SF .

1. Etapa: O processamento de dados foi realizado conforme descrito na Seção 5.1.1.
2. Etapa: Os rótulos reais do CPO foram limpos.
3. Etapa: A implementação das FR, descrita na Seção 4.4, foi aplicada ao CPO limpo para criar um banco de dados com rótulos fracos.
4. Etapa: O modelo *Sklearn-CRF* foi treinado sobre a base de supervisão fraca gerada, com aplicação da validação cruzada K-fold com cinco *folds* e com as seguintes camadas:
   - Camada de *embeddings* para transformar índices das palavras em vetores numéricos para que a rede neural aprenda representações semânticas mais significativas para as palavras
   - Camada *Convolutional Neural Network* Convi1D para extração de características e otimização da rede neural
   - Camada de *batchNormalization* para melhorar a estabilidade e a velocidade de treinamento da rede neural
   - Camada Bi-LSTM que permite que a rede neural processe as sequências de dados em duas direções simultaneamente
   - Camada de *dropout* para evitar o *overfitting*
   - Camada de *TimeDistributed* usada para aplicar uma camada densa em cada passo de tempo da sequência de entrada antes de alimentar os resultados para a camada LSTM
5. Etapa: As partições de teste utilizadas na validação do treinamento sobre a CPO foram recuperadas.
6. Etapa: Os rótulos reais dessas partições de teste foram limpos.
7. Etapa: O modelo treinado foram aplicados às partições de teste para gerar rótulos preditos.
8. Etapa: Os rótulos preditos gerados foram avaliados em relação aos rótulos originais do CPO.
9. Etapa: Os resultados foram comparados conforme descrito na Seção 5.3.

5.2.2 Experimento 2 - Aplicação Direta da Supervisão Fraca

O experimento 2 teve como objetivo comparar o desempenho de rótulos gerados exclusivamente com a aplicação de SF versus rótulos treinados usando NER tradicional no CPO. O experimento também avaliou o desempenho dos modelos NER treinados sobre a base de licitação gerada pela SF.

A etapas realizadas para conduzir esta investigação são as especificadas abaixo.

1. Etapa: Os *folds* de teste utilizados no Experimento 1, descrito na Seção 5.2.1, foram recuperados.

2. Etapa: A segunda etapa envolveu a limpeza dos rótulos reais dos *folds* de teste.
3. Etapa: O código de SF, definido na Seção 4.4, foi então aplicado às partições de teste para gerar rótulos fracos.
4. Etapa: Os rótulos fracos gerados foram avaliados comparando-os com os rótulos originais do CPO.
5. Etapa: Os resultados foram comparados e analisados conforme descrito na Seção 5.3.

5.3 Avaliação

Esta seção descreve o método de avaliação utilizado para validar se os objetivos de pesquisa elencados na Seção 1.3 foram alcançados. Os modelos de NER foram treinados com dados anotados por humanos, base ouro, e modelos treinados por SF. Dessa forma, o objetivo foi comparar os resultados alcançados com aplicação dos modelos NER sobre o CPO, com os resultados obtidos dos modelos NER sobre a base gerada pela supervisão fraca, melhor explicado na Seção 5.2. Também foi comparado os resultados gerados com aplicação da SF diretamente ao *corpus* de teste, para validar se os resultados apresentados seriam melhores.

Dentre o conjunto de dados selecionados, conforme Seção 5.1, foi aplicada a validação cruzada *K-fold* para permitir que fosse treinado 5 modelos diferentes. Em cada modelo foi utilizado uma partição de 20% para o conjunto de dados de teste e o restante para o conjunto de dados de treinamento. Dessa forma, ao utilizar a validação cruzada *K-fold* todos os dados do conjunto de dados foram utilizados tanto para treinamento quanto para teste, permitindo avaliar melhor o desempenho do modelo.

As regras da conferência *CoNLL* (Sang and De Meulder, 2003) foram seguidas para comparar os resultados, e apenas correspondências exatas foram consideradas. O que significa que só foi dado crédito a correspondências exatas ou, em outras palavras, tanto as *tags* de entidade quanto os limites tinham que estar corretos para contar como uma correspondência correta. Dessa forma, foi possível comparar a qualidade da proposta dessa pesquisa de realizar anotação baseada em SF por meio da avaliação dos resultados obtidos pela métrica *F1 Score* macro para calcular a média de acerto para todos os rótulos.

A comparação dos resultados se deu pela análise do percentual de revocação, precisão e *F1 Score* agrupados por classes. Essas métricas foram obtidas com o uso do pacote *sklearn-crfsuite.metrics* e do *framework seqeval*, conforme descrito na Seção 5.3.1

5.3.1 Framework seqeval

O *framework seqeval* [4] foi adotado para calcular a precisão, revocação e medida *F1 Score*. *Seqeval* é um *framework* em *Python* para avaliar o desempenho de modelos de rótulo de sequência. Nesta pesquisa o *Seqeval* foi configurado com o esquema do tipo *IOB2* e modelo do tipo *strict*, o que significa que as *tags IOB* deveriam corresponder na sua plenitude para serem consideradas na avaliação, conforme pode ser observado na Figura 5.2.

[4]Disponível em `https://github.com/chakki-works/seqeval`

```
>>> from seqeval.metrics import classification_report
>>> from seqeval.scheme import IOB2
>>> y_true = [['B-NP', 'I-NP', 'O']]
>>> y_pred = [['I-NP', 'I-NP', 'O']]
>>> classification_report(y_true, y_pred)
              precision    recall  f1-score   support

          NP       1.00      1.00      1.00         1

   micro avg       1.00      1.00      1.00         1
   macro avg       1.00      1.00      1.00         1
weighted avg       1.00      1.00      1.00         1

>>> classification_report(y_true, y_pred, mode='strict', scheme=IOB2)
              precision    recall  f1-score   support

          NP       0.00      0.00      0.00         1

   micro avg       0.00      0.00      0.00         1
   macro avg       0.00      0.00      0.00         1
weighted avg       0.00      0.00      0.00         1
```

Figura 5.2: Exemplo da aplicação do framework *Seqeval* Nakayama (2018)

Capítulo 6

Resultados

Esta seção apresenta os resultados dos experimentos obtidos com o treinamento do modelo de NER utilizando os algoritmos CRF e Bi-LSTM sobre o CPO e sobre a base gerada pela SF, conforme detalhado na Seção 5.2.1 e os resultados obtidos com aplicação apenas do conjunto de FR, conforme detalhado na Seção 5.2.2,

Os resultados apresentados referem-se às médias das 5 iterações usadas para validar os experimentos. As médias são exibidas agrupadas por tipo de ato de contrato e licitação e podem ser visualizados nas tabelas 6.1, 6.2, 6.3, 6.4, 6.5 e 6.6. O detalhamento dos resultados de cada partição está disponível para visualização no Apêndice C. Para facilitar a análise dos resultados dos experimentos, destacou-se em negrito o *F1 Score* que obteve o melhor desempenho.

Os resultados do experimento 1 são apresentados nas colunas: CRF, Bi-LSTM, CRF + SF e Bi-LSTM + SF, respectivamente. As colunas CRF e Bi-LSTM apresentam os resultados de um modelo tradicional de NER treinado sobre o CPO, enquanto as colunas CRF + SF e Bi-LSTM + SF apresentam os resultados dos modelos treinado sobre a base gerada pela supervisão fraca. Os resultados do experimento 2 são apresentados na coluna SF.

Entidade	CRF	CRF + SF	Bi-LSTM	Bi-LSTM + SF	SF
numero contrato	0.885	0.852	0.950	**0.966**	0.908
processo GDF	0.901	0.884	0.957	**0.958**	0.956
data assinatura contrato	0.827	0.717	0.929	**0.945**	0.778
unidade orçamentaria	0.935	0.812	0.962	**0.986**	0.826
programa trabalho	0.886	0.763	0.950	**0.975**	0.820
natureza despesa	0.913	0.868	0.950	**0.969**	0.881
nota empenho	0.843	0.827	**0.948**	0.937	0.932
valor contrato	0.834	0.765	0.886	**0.968**	0.851
Média Geral	0.885	0.819	0.950	**0.967**	0.866

Tabela 6.1: Média Extrato de Contrato

Entidade	CRF	CRF + SF	Bi-LSTM	Bi-LSTM + SF	SF
numero contrato	0.838	0.812	0.829	**0.949**	0.838
processo GDF	0.782	0.835	0.852	**0.888**	0.826
data escrito	0.852	0.827	0.901	**0.946**	0.785
Média Geral	0.838	0.827	0.852	**0.946**	0.826

Tabela 6.2: Média Aditamento de Contrato

Entidade	CRF	CRF + SF	Bi-LSTM	Bi-LSTM + SF	SF
processo GDF	0.933	0.935	0.926	**0.936**	0.935
modalidade licitacao	0.959	0.889	0.943	**0.983**	0.901
numero licitacao	0.958	0.879	0.954	**0.962**	0.884
valor estimado contratacao	0.918	0.862	0.908	**0.922**	0.857
Média Geral	0.945	0.884	0.934	**0.949**	0.892

Tabela 6.3: Média Aviso de Licitação

Entidade	CRF	CRF + SF	Bi-LSTM	Bi-LSTM + SF	SF
numero licitacao	**0.961**	0.887	0.911	0.794	0.850
modalidade licitacao	**0.955**	0.864	0.750	0.901	0.783
processo GDF	**0.935**	0.660	0.653	0.357	0.880
Média Geral	**0.955**	0.864	0.750	0.794	0.850

Tabela 6.4: Média Suspensão de Licitação de Contrato

Entidade	CRF	CRF + SF	Bi-LSTM	Bi-LSTM + SF	SF
numero licitacao	**0.766**	0.672	0.551	0.384	0.682
modalidade licitacao	0.883	0.794	0.699	**0.941**	0.940
processo GDF	0.548	0.397	0.040	0.252	**0.753**
data escrito	0.410	0.614	0.000	**0.739**	0.737
Média Geral	0.657	0.643	0.295	0.561	**0.745**

Tabela 6.5: Média Anulação e Revogação de Licitação

Entidade	CRF	CRF + SF	Bi-LSTM	Bi-LSTM + SF	SF
numero convenio	**0.675**	0.285	0.000	0.000	0.233
processo GDF	0.372	0.444	0.000	0.000	**0.897**
data assinatura convenio	**0.664**	0.000	0.000	0.000	0.115
valor convenio	**0.737**	0.663	0.000	0.000	0.711
Média Geral	**0.669**	0.364	0.000	0.000	0.472

Tabela 6.6: Média Extrato de Convênio

6.1 Análise dos resultados

A Tabela 6.1 de atos de extrato de contrato alcançou o *F1 Score* médio de 96,6% com aplicação dos modelos treinados por Bi-LSTM sobre a base da SF, contra 95% do Bi-

LSTM, que neste cenário obteve melhores resultados em relação ao CRF. Dessa forma, observa-se que os resultados alcançados com aplicação da SF superou os resultados alcançados pelos treinamentos tradicionais sobre o CPO. Já os resultados da aplicação apenas das FR, foram próximos, e em alguns casos, superiores, aos resultados alcançados pelos treinamentos tradicionais sobre o CPO, alcançando *F1 Score* médio de 95,6%.

A Tabela 6.2 de atos de aditamento contratual alcançou o *F1 Score* médio de 94,6% com aplicação dos modelos treinados por Bi-LSTM sobre a base gerada por SF, contra 85,2% do Bi-LSTM que alcançou melhores resultados em relação ao CRF. Dessa forma, observa-se que o desempenho alcançado pelos modelos treinados a partir da base de SF superou os resultados alcançados pelos treinamentos tradicionais sobre o CPO. Já os resultados da aplicação apenas das FR, foram próximos aos resultados alcançados pelos treinamentos tradicionais sobre o CPO, alcançando *F1 Score* médio de 82,6%.

A Tabela 6.3 de atos de aviso de licitação alcançou o *F1 Score* médio de 94,9% com aplicação dos modelos treinados por Bi-LSTM sobre a base gerada pela SF, contra 94,5% do CRF, que alcançou melhores resultados em relação ao Bi-LSTM. Esses resultados demonstram que a desempenho da aplicação da abordagem da SF foi superior aos resultados obtidos pelos treinamentos tradicionais sobre o CPO, e que os algoritmos tradicionais baseado em rede neural tendem a atingir resultados inferiores quando a quantidade de exemplos de treinamento é menor. Já os resultados da aplicação apenas das FR, foram próximos aos resultados alcançados pelos treinamentos tradicionais sobre o CPO, com *F1 Score* médio apresentando uma diferença de apenas 0,12%.

A Tabela 6.4 de atos de suspensão de licitação evidencia que os resultados alcançados pelos modelos treinados sobre a base da SF, foram próximos aos resultados alcançados pelos treinamentos tradicionais sobre o CPO, com *F1 Score* médio apresentando uma diferença de apenas 9,10%. Esse mesmo comportamento pode ser observado na aplicação apenas das FR, com *F1 Score* médio apresentando uma diferença de 10,5%. Pode-se observar, ainda, que os resultados dos algoritmos tradicionais baseado em rede neural tendem a atingir resultados inferiores quando a quantidade de exemplos de treinamento é menor, conforme ocorreu com os atos de aviso de licitação.

A Tabela 6.5 de atos de anulação e revogação de licitação evidencia que os resultados alcançados pela aplicação apenas das fFR, alcançou *F1 Score* médio de 74,5% contra 65,7% do CRF. E um uma diferença de 1,4% de *F1 Score* médio entre os resultados do CRF e os resultados com aplicação dos modelos treinados por CRF sobre a base gerada pela SF. Esses resultados demonstram que a desempenho alcançado com a aplicação das FR foi superior aos resultados obtidos pelos treinamentos tradicionais sobre o CPO, e que os resultados dos modelos treinados sobre a base da SF alcançaram resultados próximos aos obtidos pelos treinamentos tradicionais sobre o CPO. Da mesma forma que ocorreu nos atos anteriores, os resultados dos algoritmos tradicionais baseado em rede neural atingiram resultados inferiores que o CRF com a quantidade de exemplos de treinamento menor.

A Tabela 6.6 de atos de convênio demonstra o *F1 Score* zerado com aplicação de Bi-LSTM com SF para todas as entidades, contra resultados de *F1 Score* do Bi-LSTM também zerados e resultados do CRF com valores baixos, variando de 37,2% a 73,7%. Enquanto a aplicação apenas das FR, coluna SF, para a entidade processo GDF, obteve resultado superior ao CRF com *F1 Score* de 89,7% contra 37,2%. Esses resultados demonstram que o treinamento de modelos sobre a base gerada pela SF e os treinamentos

tradicionais sobre o CPO em *datasets* que possuem poucos exemplos ficam comprometidos em ambos os casos. Enquanto a aplicação apenas das FR SF em entidades com padrões bem definidos alcança melhores resultados mesmo com a quantidade de exemplos reduzida.

6.2 Discussão

Os resultados obtidos nesta análise destacam a importância e o potencial das abordagens semi-supervisionadas, especialmente a técnica de SF, na geração de dados anotados e no treinamento de modelos de NER em cenários em que há escassez de exemplos anotados manualmente em português.

Os experimentos revelam que a aplicação dos modelos treinados sobre a base gerada pela SF superou consistentemente os resultados dos treinamentos tradicionais realizados sobre o CPO em diversas categorias de atos, como os de extrato de contrato, aditamento contratual e aviso de licitação. Esses resultados evidenciam a eficácia da SF na geração de dados anotados de qualidade, possibilitando o desenvolvimento de modelos de NER com desempenho comparável ou superior aos obtidos com dados anotados manualmente. No entanto, os resultados foram baixos para entidades com poucos exemplos ou com padrões de escrita pouco definidos nos textos analisados.

Em relação a aplicação apenas das FR na base de licitação e contratos, os resultados foram semelhantes ou até superiores aos modelos treinados no CPO, especialmente quando as entidades anotadas apresentavam um padrão bem definido. Isso sugere que abordagem da SF pode ser uma alternativa viável e eficaz para a rotulação automática de dados em cenários onde a disponibilidade de exemplos anotados manualmente é limitada.

Acredita-se que a eficácia da aplicação apenas do SF está relacionada a qualidade das regras definidas a partir do conhecimento heurístico e do domínio em que o *corpus* foi utilizado. Assim como, o fato da quantidade de exemplos rotulados influenciar mais entre a diferença entre os resultados dos modelos de aprendizagem de máquina em detrimento a aplicação da SF.

Em resumo, é possível observar que o desempenho alcançado pelos modelos treinados sobre a base da SF e o desempenho alcançado apenas com a aplicação das FR, demonstram uma importante contribuição para a geração automatizada de bases anotadas, além de apresentarem um custo de construção muito menor que o custo para realizar a rotulação manual.

Capítulo 7

Conclusões

Este estudo foi impulsionado pela questão central de como gerar modelos automatizados e precisos para a tarefa de NER, especialmente em cenários com poucos dados anotados disponíveis em português e características contextuais específicas que demandam anotação especializada. A hipótese investigada foi se as técnicas de SF poderiam contribuir para a rápida geração de dados anotados e para a criação de modelos com qualidade comparável aos modelos anotados por humanos.

A direção da pesquisa, ancorada nessa questão central, proporcionou um foco claro e uma estrutura sólida para os experimentos. A investigação da eficácia das técnicas de SF na geração de dados anotados de qualidade culminou no objetivo geral de criar um corpus de Licitação e Contratação Pública anotado automaticamente, utilizando métodos de aprendizado fracamente supervisionado, e comparar o desempenho de modelos de aprendizagem de máquina treinados sobre a base CPO com modelos treinados na base criada esses métodos. O diferencial deste estudo reside no fato de que, apesar da variedade de abordagens para aplicar SF na tarefa de NER, não foram encontrados trabalhos com foco na aplicação de SF utilizando FR de correspondência por palavra, conhecimento heurístico (regras) e aprendizado de máquina para geração de bases anotadas no domínio brasileiro.

Para orientar esta pesquisa, foram definidos objetivos específicos, incluindo a investigação de métodos de geração de entidades nomeadas usando a abordagem de SF, a seleção de entidades de licitação e contrato, a descrição dos padrões ou características das entidades selecionadas e a implementação das FR para criar uma base de dados anotados com a abordagem de SF.

Os experimentos realizados para validar a hipótese incluíram o treinamento de NER tanto no CPO quanto na base gerada pela SF, juntamente com uma análise comparativa de desempenho entre os rótulos gerados exclusivamente pela aplicação da SF. Esses experimentos foram conduzidos com rigor metodológico, incorporando otimizações nos hiper parâmetros dos modelos, melhorias nas FR e uma abordagem cuidadosa na seleção e adaptação das técnicas de SF. Essas ações corretivas foram cruciais para garantir a robustez e a qualidade dos resultados obtidos.

Os resultados confirmaram a hipótese, demonstrando que os modelos treinados na base gerada pela SF superaram consistentemente os modelos treinados na CPO em diversas categorias de atos. Esses resultados ressaltam a eficácia da abordagem de SF na geração de

dados anotados automaticamente e no treinamento de modelos de NER com desempenho comparável ou até superior aos obtidos com dados anotados manualmente.

Dessa forma, é possível gerar modelos automatizados e precisos para a tarefa de NER, mesmo em contextos com poucos dados anotados disponíveis em português e com características contextuais específicas, corroborando com o viés da literatura, o qual evidencia que a adoção dessa abordagem é promissora na automatização da geração de bases anotadas (Lison et al., 2021). Nesse sentido, pode-se afirmar que o uso de métodos fracamente supervisionados pode beneficiar a construção de aplicações que requeiram uma etapa de aprendizado de máquina baseado em uma tarefa de NER. Isso confirma positivamente a hipótese estabelecida, além de apresentar um resultado relevante ao disponibilizar um conjunto de dados anotados de licitação e contratos que podem ser utilizados em aplicações que exigem o treinamento de modelos de NER no domínio da língua portuguesa. É importante destacar que o custo e o tempo necessários para criar FR são significativamente menores do que o custo da anotação humana, conforme evidenciado na Seção 4.1.

Em síntese, os resultados deste estudo reforçam a importância das abordagens semi-supervisionadas, particularmente a abordagem de programação de dados da SF, como uma ferramenta promissora para a geração de dados anotados e o treinamento de modelos de NER em contextos onde a anotação manual é impraticável ou inviável. Tais abordagens têm o potencial de acelerar significativamente o desenvolvimento de sistemas de processamento de linguagem natural em português, abrindo novas perspectivas para aplicações em diversos domínios e oferecendo resultados de alta qualidade com redução de custos e tempo.

7.1 Trabalhos futuros

Apesar dos resultados promissores, a investigação precisa ser aprofundada. Como trabalhos futuros, pretende-se expandir as FR para abarcar as demais entidades e assim prover uma base deados de atos de licitação e contrato com uma quantidade maior de entidades.

Embora os experimentos tenham apresentado bons resultados, a aplicação de outras técnicas, incluindo a transferência de aprendizado e aprendizado ativo, pode resultar em investigações interessantes.

Considerando os resultados alcançados até o momento, pode-se afirmar que a abordagem da SF combinada com o uso dos algoritmos de NER apresentou uma boa capacidade de predição e indicação dos rótulos automatizadamente. No entanto, seria recomendável conduzir novos testes em uma base de licitação e contratos melhor qualificada. Certamente, esse caminho requer a parceria com outros órgãos para acesso à base completa de licitações e contratos do sistema de compras do governo federal. O que durante o tempo de execução de um mestrado acadêmico, e principalmente em época de pandemia, não se mostrou viável.

A fim de tornar o modelo mais factível de adoção, sugere que seja criado um *framework* para permitir o reuso das FR em diferentes contextos.

Referências

Albuquerque, H. O., Costa, R., Silvestre, G., Souza, E., da Silva, N. F. F., Vitório, D., Moriyama, G., Martins, L., Soezima, L., Nunes, A., Siqueira, F., Tarrega, J. P., Beinotti, J. V., Dias, M., Silva, M., Gardini, M., Silva, V., de Carvalho, A. C. P. L. F., and Oliveira, A. L. I. (2022). Ulyssesner-br: A corpus of brazilian legislative documents for named entity recognition. In *Computational Processing of the Portuguese Language*. Springer International Publishing. 24, 25, 27

Bach, S. H., Rodriguez, D., Liu, Y., Luo, C., Shao, H., Xia, C., Sen, S., Ratner, A., Hancock, B., Alborzi, H., Kuchhal, R., Ré, C., and Malkin, R. (2019). Snorkel drybell: A case study in deploying weak supervision at industrial scale. In *Proceedings of the 2019 International Conference on Management of Data*, page 362–375. Association for Computing Machinery. 1, 17, 31

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. 18

Chen, Li-Ming, B.-X. and Ding, Z.-Y. (2022). Multiple weak supervision for short text classification. *Applied Intelligence*, 52(8):9101–9116. 24

Chiu, J. P. C. and Nichols, E. (2015). Named entity recognition with bidirectional lstm-cnns. cite arxiv:1511.08308. ix, 2, 14

Deng, K., Wang, D., and Liu, J. (2017). *Weakly-supervised named entity extraction using word representations*, volume 10179 LNCS of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Bioinformatics. 82

Gao, N., Zhu, Z., Weng, Z., Chen, G., and Zhang, M. (2020). A supervised named entity recognition method based on pattern matching and semantic verification. *Journal of Internet Technology*, 21(7):1917–1928. 82

Guo, J., Xu, G., Cheng, X., and Li, H. (2009). Named entity recognition in query. In *SIGIR*, page 267–274, New York, NY, USA. Association for Computing Machinery. 78, 81, 82

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780. 13

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear. 31

Jo, T. (2021). *Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning*. Springer. 15, 16

Jurafsky, D. and Martin, J. H. (2009). *Speech and Language Processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson international edition. Pearson Prentice Hall/Pearson education international. 9, 10

Knofczynski, J., Durairajan, R., and Willinger, W. (2022). Arise: A multitask weak supervision framework for network measurements. *IEEE Journal on Selected Areas in Communications*, 40(8):2456–2473. 23, 24

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2015). DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195. 18

Li, J., Ding, H., Shang, J., McAuley, J., and Feng, Z. (2021). Weakly supervised named entity tagging with learnable logical rules. 22, 24, 25, 26

Li, Z., Chao, J., Zhang, M., Chen, W., Zhang, M., and Fu, G. (2017). Coupled pos tagging on heterogeneous annotations. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 25(3):557–571. 82

Lison, P., Barnes, J., and Hubin, A. (2021). skweak: Weak supervision made easy for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Association for Computational Linguistics. 18, 20, 28, 31, 34, 38, 40, 54, 81

Lison, P., Hubin, A., Barnes, J., and Touileb, S. (2020). Named entity recognition without labelled data: A weak supervision approach. ix, 2, 3, 16, 17, 20, 21, 23, 24, 25, 26, 81

Luz de Araujo, P. H., de Campos, T. E., de Oliveira, R. R. R., Stauffer, M., Couto, S., and Bermejo, P. (2018). LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), pages 313–323, Canela, RS, Brazil. Springer. 2, 24, 25, 27

Mariano, A. M. Rocha, M. (2017). Revisão da literatura: Apresentação de uma abordagem integradora. *Proceedings of XXVI Congreso Internacional de la Academia Europea de Dirección y Economía de la Empresa (AEDEM)*, 26:430–438. ix, 22, 73, 74

Martins, J. S., Lenz, M. L., Silva, M. B. F. d., Oliveira, R. A. d., Pichetti, R. F., Mariano, D. C. B., Martins, J. V., Rodrigues, S. M. A. F., and Bezerra, W. R. (2020). *Processamentos de Linguagem Natural*. Porto Alegre, RS : Sagah, 1 edition. 9

Nakayama, H. (2018). seqeval: A python frameworzk for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval. ix, 48

Nguyen, A. T., Wallace, B., Li, J. J., Nenkova, A., and Lease, M. (2017). Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 299–309, Vancouver, Canada. Association for Computational Linguistics. 20

Nothman, J., Ringland, N., Radford, W., Murphy, T., and Curran, J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175. 19

Rabiner, L. and Juang, B. (1986). An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16. 20

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282. 17, 18, 20, 40

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition. 8

Safranchik, E., Luo, S., and Bach, S. (2020). Weakly supervised sequence tagging from noisy rules. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5570–5578. 17, 18, 23, 24, 25, 26

Sang, E. F. T. K. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. 2, 10, 18, 31, 47

Schmidt, R. M. (2019). Recurrent neural networks (rnns): A gentle introduction and overview. ix, 12, 13, 30

Sedova, A., Stephan, A., Speranskaya, M., and Roth, B. (2021). Knodle: Modular weakly supervised learning with pytorch. *CoRR*, abs/2104.11557. 31

Shang, J., Liu, L., Ren, X., Gu, X., Ren, T., and Han, J. (2018). Learning named entity tagger using domain-specific dictionary. 23, 24, 26

Sugiyama, M., Bao, H., Ishida, T., Lu, N. L., and Sakai, T. (2022). *Machine Learning from Weak Supervision: An Empirical Risk Minimization Approach*. Adaptive Computation and Machine Learning Series. The MIT Press. 16

Sutton, C. and McCallum, A. (2012). An introduction to conditional random fields. *Found. Trends Mach. Learn.*, 4(4):267–373. 10, 11, 12

Tok, W. H., Bahree, A., and Filipi, S. (2022). *Practical Weak Supervision-Doing More with Less Data*. O'Reilly Media, Inc., 2 edition. 1, 2, 16, 17, 18, 20

Wang, F., Wu, W., Li, Z., and Zhou, M. (2017). Named entity disambiguation for questions in community question answering. *Knowledge-Based Systems*, 126:68–77. 82

Wang, X., Zhang, Y., Li, Q., Wu, C. H., and Han, J. (2019). Penner: Pattern-enhanced nested named entity recognition in biomedical literature. In elsevier, editor, *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, pages 540–547. 82

Weischedel, Ralph, e. a. (2013). Ontonotes release 5.0 ldc2013t19. 2, 18, 31

Wick, M. (2015). Geonames ontology. 19

Wißler, L., Almashraee, M., Díaz, D. M., and Paschke, A. (2014). The gold standard in corpus annotation. In *IEEE GSC*. IEEE. 2

Xu, G., Yang, S., and Li, H. (2009). Named entity mining from click-through data using weakly supervised latent dirichlet allocation. In IV, J. F. E., Fogelman-Soulié, F., Flach, P. A., and Zaki, M. J., editors, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, pages 1365–1374. ACM. 81, 82

Zhao, Q., Wang, D., Xu, S., Zhang, X., and Wang, X. (2020). A weakly supervised chinese medical named entity recognition method. *Harbin Gongcheng Daxue Xuebao/Journal of Harbin Engineering University*, 41(3). 82

Zhou, Z.-H. (2017). A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53. 1, 2

Zupic, I. Čater, T. (2015). Bibliometric methods in management and organization. organizational research methods. *Human and Social Management*, 18:429–472. 81

Apêndice A

Tabela dos Atos e Entidades

Tabela A.1: Lista dos Atos e entidades

Aviso de Suspensão	**Aviso de Revogação/Anulação**	**Extrato de Contrato**	**Extrato de Convênio**	**Extrato de aditamento**	**Aviso de Abertura**
numero licitacao	numero licitacao	cnpj entidade contratada	entidade convenente	numero contrato	numero licitacao
nome responsavel	identificacao ocorrencia	cnpj orgao contratante	data assinatura convenio	orgao contratante	tipo objeto
modalidade licitacao	modalidade licitacao	codigo siggo	vigencia convenio	numero termo aditivo	modalidade licitacao
orgao licitante	orgao licitante	data assinatura contrato	objeto convenio	objeto aditamento contratual	orgao licitante
objeto licitacao	nome responsavel	entidade contratada	processo gdf	processo gdf	objeto licitacao
processo gdf	processo gdf	fonte recurso	numero convenio	data escrito	processo gdf
decisao tcdf	data escrito	natureza despesa	cnpj entidade convenente	nome responsavel	data abertura licitacao
		nome responsavel	valor convenio	codigo siggo	nome responsavel
		nota empenho	orgao concedente	numero convenio	sistema compras
		numero contrato	natureza despesa		valor estimado contratacao
		objeto contrato	programa trabalho		codigo licitacao sistema compras
		orgao contratante	nota empenho		
		processo gdf	fonte recurso		
		programa trabalho	unidade orcamentaria		
		unidade orcamentaria	objeto contrato		
		valor contrato	cnpj orgao concedente		
		vigencia contrato	vigencia contrato		
			data assinatura contrato		
			nome responsavel		

Apêndice B

Tabelas descritivas do corpus padrão ouro (CPO)

O corpus padrão ouro contemplou 4090 atos de licitação e contrato e 43.733 entidades, e é composto por cinco tipos de atos: Extrato de Contrato, Aditamento de Contrato, Aviso de Licitação, Suspensão de Licitação, Anulação e Revogação de Licitação e Extrato de Convênio. A lista das entidades selecionadas e quantidade de exemplos por entidade estão descritos nas tabelas B.1, B.2, B.3, B.4, B.5, B.6.

Tabela B.1: Entidades do Ato Extrato de Contrato

Entidades	Qtde
orgao_contratante	2225
numero_contrato	1900
entidade_contratada	1769
processo_gdf	1758
objeto_contrato	1726
vigencia_contrato	1664
valor_contrato	1586
fonte_recurso	1440
programa_trabalho	1409
nota_empenho	1392
data_assinatura_contrato	1287
natureza_despesa	1188
unidade_orcamentaria	1108
cnpj_entidade_contratada	485
nome_responsavel	185
codigo_siggo	175
cnpj_orgao_contratante	135
Total	23166

Tabela B.2: Entidades do Ato Aditamento de Contrato

Entidades	Qtde
numero_contrato	2153
orgao_contratante	1830
numero_termo_aditivo	1632
objeto_aditamento_contratual	1536
processo_gdf	1487
data_escrito	1429
codigo_siggo	130
nome_responsavel	96
Total	10293

Tabela B.3: Entidades do Ato Aviso de Licitação

Entidades	Qtde
numero_licitacao	712
modalidade_licitacao	703
sistema_compras	696
tipo_objeto	654
objeto_licitacao	637
data_abertura_licitacao	629
nome_responsavel	629
processo_gdf	616
orgao_licitante	560
valor_estimado_contratacao	509
codigo_licitacao_sistema_compras	445
Total	6790

Tabela B.4: Entidades do Ato Suspensão de Licitação

Entidades	Qtde
numero_licitacao	97
modalidade_licitacao	95
nome_responsavel	81
orgao_licitante	78
objeto_licitacao	71
processo_gdf	63
decisao_tcdf	18
Total	503

Tabela B.5: Entidades do Ato Anulação e Revogação de Licitação

Entidades	Qtde
numero_licitacao	66
modalidade_licitacao	65
identificacao_ocorrencia	55
nome_responsavel	51
processo_gdf	46
orgao_licitante	44
data_escrito	18
Total	345

Tabela B.6: Entidades do Ato Extrato de Convênio

Entidades	Qtde
entidade_convenente	39
objeto_convenio	31
data_assinatura_convenio	30
vigencia_convenio	28
processo_gdf	28
numero_convenio	25
orgao_concedente	23
valor_convenio	15
cnpj_entidade_convenente	13
programa_trabalho	9
natureza_despesa	9
fonte_recurso	7
nota_empenho	6
unidade_orcamentaria	5
cnpj_orgao_concedente	5
nome_responsavel	4
objeto_contrato	2
vigencia_contrato	1
Total	280

Apêndice C

Tabelas de Resultados Por Interações

Para cada ato de licitação e contrato foram extraídos os resultados das cinco interações executadas pelos algoritmos, a média das interações e o desvio padrão.

As tabelas C.1, C.6, C.11, C.16, C.21 e C.26, exibem resultado do treinamento de NER tradicional realizado utilizando o algoritmo CRF. Enquanto que as tabelas C.2, C.7, C.12, C.17, C.22 e C.27 exibem resultado do treinamento NER do CRF sobre a base de dados gerada pela SF.

Já as tabelas C.3, C.8, C.13, C.18, C.23 e C.28, exibem resultado do treinamento de NER tradicional realizado utilizando o algoritmo Bi-LSTM. Enquanto que as tabelas C.4, C.9, C.14, C.19, C.24 e C.29 exibem resultado do treinamento NER do Bi-LSTM sobre a base de dados gerada pela SF.

As tabelas C.5, C.10, C.15, C.20, C.25 e C.30 exibem o resultado da aplicação direta da SF.

C.1 Resultados Extrato de Contrato

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero contrato	0.879	0.882	0.884	0.902	0.875	**0.885**	0.010
processo gdf	0.901	0.913	0.895	0.906	0.888	**0.901**	0.010
data assinatura contrato	0.855	0.787	0.839	0.815	0.837	**0.827**	0.026
unidade orcamentaria	0.935	0.913	0.957	0.935	0.933	**0.935**	0.016
programa trabalho	0.866	0.907	0.890	0.877	0.890	**0.886**	0.015
natureza despesa	0.882	0.955	0.934	0.899	0.897	**0.913**	0.030
nota empenho	0.845	0.834	0.825	0.867	0.846	**0.843**	0.016
valor contrato	0.808	0.789	0.870	0.868	0.833	**0.834**	0.036

Tabela C.1: Extrato de Contrato CRF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero contrato	0.846	0.855	0.838	0.873	0.847	**0.852**	0.014
processo gdf	0.879	0.895	0.883	0.880	0.881	**0.884**	0.007
data assinatura contrato	0.681	0.706	0.717	0.752	0.730	**0.717**	0.026
unidade orcamentaria	0.783	0.806	0.847	0.829	0.797	**0.812**	0.025
programa trabalho	0.733	0.789	0.780	0.769	0.744	**0.763**	0.024
natureza despesa	0.877	0.897	0.880	0.845	0.843	**0.868**	0.024
nota empenho	0.828	0.843	0.785	0.844	0.837	**0.827**	0.025
valor contrato	0.737	0.719	0.779	0.795	0.797	**0.765**	0.035

Tabela C.2: Extrato de Contrato CRF + SF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero contrato	0.965	0.941	0.945	0.954	0.946	**0.950**	0.01
processo gdf	0.964	0.954	0.954	0.974	0.939	**0.957**	0.013
data assinatura contrato	0.927	0.914	0.952	0.925	0.927	**0.929**	0.014
unidade orcamentaria	0.962	0.952	0.975	0.943	0.978	**0.962**	0.015
programa trabalho	0.914	0.980	0.937	0.967	0.951	**0.950**	0.026
natureza despesa	0.946	0.955	0.969	0.953	0.924	**0.950**	0.017
nota empenho	0.924	0.954	0.956	0.958	0.947	**0.948**	0.014
valor contrato	0.884	0.868	0.902	0.906	0.870	**0.886**	0.017

Tabela C.3: Extrato de Contrato Bi-LSTM

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero contrato	0.961	0.975	0.964	0.966	0.967	**0.966**	0.005
processo gdf	0.941	0.958	0.980	0.955	0.949	**0.957**	0.015
data assinatura contrato	0.942	0.955	0.946	0.951	0.928	**0.945**	0.011
unidade orcamentaria	0.991	0.984	0.993	0.983	0.980	**0.986**	0.006
programa trabalho	0.966	0.985	0.979	0.982	0.961	**0.975**	0.011
natureza despesa	0.966	0.971	0.986	0.957	0.967	**0.969**	0.010
nota empenho	0.935	0.971	0.890	0.956	0.934	**0.937**	0.030
valor contrato	0.971	0.979	0.975	0.954	0.961	**0.968**	0.011

Tabela C.4: Extrato de Contrato Bi-LSTM + SF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero contrato	0.918	0.906	0.909	0.912	0.897	**0.908**	0.008
processo gdf	0.968	0.945	0.945	0.963	0.961	**0.956**	0.011
data assinatura contrato	0.758	0.768	0.773	0.805	0.783	**0.778**	0.018
unidade orcamentaria	0.813	0.813	0.835	0.858	0.810	**0.826**	0.021
programa trabalho	0.833	0.846	0.815	0.798	0.807	**0.820**	0.019
natureza despesa	0.904	0.887	0.867	0.862	0.883	**0.881**	0.017
nota empenho	0.912	0.953	0.926	0.930	0.938	**0.932**	0.015
valor contrato	0.812	0.797	0.805	0.785	0.836	**0.851**	0.019

Tabela C.5: Extrato de Contrato SF

C.2 Resultados Aditamento de Contrato

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero contrato	0.818	0.853	0.852	0.840	0.828	**0.838**	0.015
processo gdf	0.788	0.794	0.780	0.782	0.768	**0.782**	0.010
data escrito	0.880	0.832	0.857	0.836	0.858	**0.852**	0.019

Tabela C.6: Aditamento de Contrato CRF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero contrato	0.815	0.808	0.810	0.801	0.827	**0.812**	0.010
processo gdf	0.848	0.856	0.846	0.803	0.825	**0.835**	0.021
data escrito	0.841	0.816	0.814	0.817	0.847	**0.827**	0.016

Tabela C.7: Aditamento de Contrato CRF + SF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero contrato	0.845	0.822	0.816	0.834	0.829	**0.829**	0.011
processo gdf	0.854	0.865	0.886	0.860	0.795	**0.852**	0.034
data escrito	0.878	0.918	0.900	0.893	0.914	**0.901**	0.016

Tabela C.8: Aditamento de Contrato Bi-LSTM

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero contrato	0.948	0.953	0.941	0.953	0.949	**0.949**	0.005
processo gdf	0.919	0.784	0.932	0.860	0.942	**0.888**	0.066
data escrito	0.967	0.945	0.958	0.924	0.933	**0.946**	0.017

Tabela C.9: Aditamento de Contrato Bi-LSTM + SF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero contrato	0.819	0.831	0.824	0.847	0.864	**0.838**	0.018
processo gdf	0.891	0.891	0.881	0.872	0.898	**0.826**	0.010
data escrito	0.769	0.799	0.775	0.786	0.794	**0.785**	0.012

Tabela C.10: Aditamento de Contrato SF

C.3 Resultados Aviso de Licitação

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
processo gdf	0.944	0.931	0.897	0.946	0.948	**0.933**	0.021
modalidade licitacao	0.968	0.958	0.965	0.949	0.956	**0.959**	0.008
numero licitacao	0.972	0.970	0.961	0.956	0.930	**0.958**	0.017
valor estimado contratacao	0.915	0.924	0.944	0.865	0.943	**0.918**	0.032

Tabela C.11: Aviso de Licitação CRF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
processo gdf	0.928	0.934	0.924	0.943	0.945	**0.935**	0.009
modalidade licitacao	0.883	0.905	0.901	0.842	0.917	**0.889**	0.029
numero licitacao	0.870	0.904	0.877	0.856	0.889	**0.879**	0.018
valor estimado contratacao	0.804	0.883	0.901	0.846	0.878	**0.862**	0.038

Tabela C.12: Aviso de Licitação CRF + SF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
processo gdf	0.899	0.928	0.950	0.936	0.915	**0.926**	0.019
modalidade licitacao	0.954	0.924	0.951	0.922	0.967	**0.943**	0.020
numero licitacao	0.966	0.966	0.964	0.930	0.945	**0.954**	0.016
valor estimado contratacao	0.895	0.933	0.921	0.853	0.942	**0.908**	0.036

Tabela C.13: Aviso de Licitação Bi-LSTM

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
processo gdf	0.930	0.980	0.913	0.917	0.940	**0.936**	0.027
modalidade licitacao	0.991	0.987	0.977	0.979	0.980	**0.983**	0.006
numero licitacao	0.922	0.967	0.973	0.974	0.972	**0.962**	0.022
valor estimado contratacao	0.956	0.783	0.948	0.950	0.975	**0.922**	0.079

Tabela C.14: Aviso de Licitação Bi-LSTM + SF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
processo gdf	0.964	0.954	0.955	0.924	0.955	**0.935**	0.015
modalidade licitacao	0.924	0.886	0.924	0.881	0.890	**0.901**	0.021
numero licitacao	0.513	0.000	0.891	0.014	0.000	**0.284**	0.405
valor estimado contratacao	0.828	0.878	0.888	0.833	0.857	**0.857**	0.026

Tabela C.15: Aviso de Licitação SF

C.4 Resultados Suspensão de Licitação de Contrato

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero licitacao	0.972	0.978	0.964	0.959	0.934	**0.961**	0.017
modalidade licitacao	0.964	0.950	0.965	0.945	0.953	**0.955**	0.009
processo gdf	0.934	0.919	0.921	0.958	0.943	**0.935**	0.016

Tabela C.16: Suspensão de Licitação de Contrato CRF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero licitacao	0.877	0.905	0.908	0.885	0.858	**0.887**	0.021
modalidade licitacao	0.862	0.863	0.887	0.827	0.880	**0.864**	0.023
processo gdf	0.656	0.717	0.597	0.655	0.674	**0.660**	0.043

Tabela C.17: Suspensão de Licitação de Contrato CRF + SF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero licitacao	0.882	1.000	0.968	0.923	0.780	**0.911**	0.085
modalidade licitacao	0.706	0.944	0.788	0.829	0.485	**0.750**	0.172
processo gdf	0.727	0.560	0.800	0.545	0.632	**0.653**	0.109

Tabela C.18: Suspensão de Licitação de Contrato Bi-LSTM

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero licitacao	0.444	0.971	0.963	0.750	0.842	**0.794**	0.216
modalidade licitacao	0.923	1.000	0.971	0.979	0.632	**0.901**	0.153
processo gdf	0.000	0.476	0.240	0.400	0.667	**0.357**	0.252

Tabela C.19: Suspensão de Licitação de Contrato Bi-LSTM + SF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero licitacao	0.841	0.893	0.860	0.831	0.826	**0.850**	0.037
modalidade licitacao	0.800	0.833	0.800	0.833	0.647	**0.783**	0.078
processo gdf	0.750	0.846	0.889	1.000	0.917	**0.880**	0.092

Tabela C.20: Suspensão de Licitação de Contrato SF

C.5 Resultados Anulação e Revogação de Licitação

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero licitacao	0.846	0.815	0.690	0.783	0.696	**0.766**	0.07
modalidade licitacao	1.000	0.880	0.846	0.833	0.857	**0.883**	0.067
processo gdf	0.462	0.615	0.333	0.615	0.714	**0.548**	0.150
data escrito	0.000	0.750	0.444	0.857	0.000	**0.410**	0.404

Tabela C.21: Anulação e Revogação de Licitação CRF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero licitacao	0.720	0.667	0.857	0.545	0.571	**0.672**	0.125
modalidade licitacao	0.800	0.833	0.696	0.783	0.857	**0.794**	0.062
processo gdf	0.200	0.500	0.000	0.714	0.571	**0.397**	0.291
data escrito	0.286	0.727	0.769	0.889	0.400	**0.614**	0.258

Tabela C.22: Anulação e Revogação de Licitação CRF + SF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero licitacao	0.667	0.316	0.737	0.444	0.593	**0.551**	0.171
modalidade licitacao	0.696	0.700	0.600	0.700	0.800	**0.699**	0.071
processo gdf	0.200	0.000	0.000	0.000	0.000	**0.040**	0.089
data escrito	0.000	0.000	0.000	0.000	0.000	**0.000**	0.000

Tabela C.23: Anulação e Revogação de Licitação Bi-LSTM

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero licitacao	0.700	0.167	0.667	0.222	0.167	**0.384**	0.274
modalidade licitacao	0.909	0.952	1.000	0.900	0.941	**0.941**	0.040
processo gdf	0.400	0.308	0.154	0.000	0.400	**0.252**	0.173
data escrito	0.909	0.727	0.444	0.889	0.727	**0.739**	0.186

Tabela C.24: Anulação e Revogação de Licitação Bi-LSTM + SF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero licitacao	0.700	0.612	0.842	0.645	0.612	**0.682**	0.215
modalidade licitacao	0.909	0.952	1.000	0.900	0.941	**0.941**	0.040
processo gdf	0.400	0.308	0.154	0.000	0.400	**0.252**	0.173
data escrito	0.909	0.727	0.444	0.889	0.727	**0.739**	0.186

Tabela C.25: Anulação e Revogação de Licitação SF

C.6 Resultados Anulação e Revogação de Licitação

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero convenio	0.909	0.444	0.500	0.857	0.667	**0.675**	0.207
processo gdf	0.000	0.571	0.400	0.000	0.889	**0.372**	0.382
data assinatura convenio	0.800	0.667	0.615	0.571	0.667	**0.664**	0.086
valor convenio	0.800	0.750	0.800	1.000	0.333	**0.737**	0.245

Tabela C.26: Extrato de Convênio CRF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero convenio	0.444	0.286	0.000	0.333	0.364	**0.285**	0.170
processo gdf	0.500	0.571	0.400	0.000	0.750	**0.444**	0.279
data assinatura convenio	0.000	0.000	0.000	0.000	0.000	**0.000**	0.000
valor convenio	1.000	0.750	0.500	0.667	0.400	**0.663**	0.233

Tabela C.27: Extrato de Convênio CRF + SF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero convenio	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0
processo gdf	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0
data assinatura convenio	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0
valor convenio	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0

Tabela C.28: Extrato de Convênio Bi-LSTM

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero convenio	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0
processo gdf	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0
data assinatura convenio	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0
valor convenio	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0

Tabela C.29: Extrato de Convênio Bi-LSTM + SF

Entidade	Fold0	Fold1	Fold2	Fold3	Fold4	Média	DP
numero convenio	0.500	0.000	0.000	0.667	0.000	**0.233**	0.325
processo gdf	1.000	0.923	0.923	0.727	0.909	**0.897**	0.101
data assinatura convenio	0.000	0.200	0.222	0.154	0.000	**0.115**	0.108
valor convenio	0.400	0.500	0.800	1.000	0.857	**0.711**	0.252

Tabela C.30: Extrato de Convênio SF

Anexo A

Protocolo TEMAC

A Revisão Sistemática (RS) é um método para pesquisa bibliográfica e foi realizada por meio da Teoria do Enfoque Meta Analítico Consolidada (TEMAC) Mariano (2017). O TEMAC está dividido em 3 etapas: preparação da pesquisa, apresentação e interrelação dos dados e detalhamento, modelo integrador e validação por evidências.

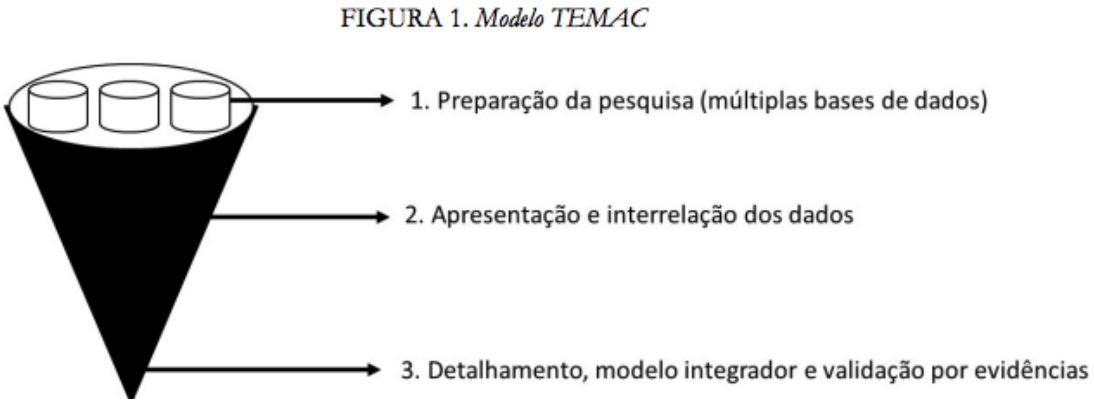


Figura A.1: Etapas TEMAC - Retirado Mariano (2017)

1. Preparação da Pesquisa: etapa que consiste na busca da palavra chave da pesquisa e na delimitação das fronteira a ser pesquisa. Para se realizar esta etapa podem ser realizadas perguntas como:

   - Qual o descritor ou palavra de pesquisa?
   - Qual o campo espaço/tempo da pesquisa?
   - Quais as bases de dados serão utilizadas?
   - Quais áreas de conhecimento serão utilizadas?

   Para se encontrar o descritor, um caminho é pesquisar o tema, identificar um artigo e selecionar suas palavras-chaves da seção resumo, e escolher a que mais se aplica ao contexto.

   As bases de dados sugeridas pelo TEMAC são: *Web of Science*, *Google* Acadêmico e *Scopus*.

2. Apresentação e interrelação dos dados e detalhamento: etapa que se realiza a análise de relação entre os registros através dos seguintes critérios: - Revistas mais relevantes - Revistas que mais publicam - Evolução do tema ano a ano - Documentos mais citados - Autores que mais publicaram versus os que mais foram citados - Países que mais publicaram - Conferências que mais contribuiram - Universidades que mais contribuiram - Agenciais que mais financiaram - Áreas que mais publicaram - Frequência de palavras chaves
3. Modelo integrador e validação por evidências: etapa onde são necessárias análises mais profundas que permitam compreender melhor o tema, como selecionar aqueles autores que não podem faltar na revisão, as principais abordagens, linhas de pesquisa, validação via evidências e entrega do modelo integrador por meio da comparação dos resultados das diferentes fontes. Por exemplo, os autores que publicaram o assunto mais de 2 vezes, o país que possui maior domínio sobre o assunto, ano da publicação, geralmente os mais recentes, etc.

A.1 Revisão através da Teoria do Enfoque Meta Analítico Consolidada (TEMAC)

A.1.1 Etapa 1: Preparação da Pesquisa

Para a construção da palavra chave de pesquisa foram pesquisadas palavras em inglês que refletissem o tema, envolvendo expressões relacionadas a supervisão fraca, sendo assim foi formado o termo: *"Weakly Supervised" OR "Named Entity" or "Weak supervision" or "Labelled Data" or "Named Entity Recognition"*. Foram escolhidas para a pesquisa as bases de dados *Web of Science* e *Scopus* por serem base de dados consolidadas e de maior qualidade Mariano (2017). A busca nas bases foi realizada em 15 de agosto de 2021. Na *Web of Science* os resultados foram filtrados pela categoria: *Computer Science Artificial Intelligence*, retornando 2867 resultados. Já na*Scopus* a pesquisa foi filtrada na área *Computer Science*, retornando 18655 resultados.
Avaliando o campo de pesquisa foi observado que a quantidade de documentos estava muito elevada, nesse sentido foi aplicado o filtro para contemplar apenas artigos abertos. Dessa forma, na *Web of Science* foram encontrando 892 resultados. Já na *Scopus* a pesquisa encontrou 5867 resultados. Após análise da palavra chave utilizada surgiu a hipótese de ampliar a *String* de consulta para:

> *"weak supervision" or "weak supervision frameworks" or "weak supervision labelling functions" or "weak supervision model" or "weak supervision approach" or "weak supervision sources" or "weak supervision to label" or "weakly supervised classification" or "weakly Supervised" or "weak supervision rules" or "weakly Supervised Named Entity" or "weak supervision NLP" or "weakly supervised machine learning" or "weakly supervised learning" or "weakly supervised modelling" or "weakly supervised model" or "weakly supervised training" or "weakly supervised data" or "weakly supervised algorithms" or "weakly supervised sequence" or "weakly supervised methods" or "weakly supervised setting" or "weakly Supervised Named Entity" or "weakly supervised NER" or "weakly supervised models" or "weakly supervise sequence".*

Com isso, reduziu a quantidade de artigos para 1579 na *Web of Science*, após filtrar pela categoria *Computer Science Artificial Intelligence* e 3519 na *Scopus*, após filtrar pela categoria *Computer Science*. Porém, a quantidade de artigos ainda se mostrou bastante elevada. Observou-se, ainda, que utilizando apenas as *strings "weak supervision" or "weakly Supervised" or "weakly supervise"* os resultados coincidiam com os da *string* anterior. Nesse sentido, optou-então por concatenar *("weak supervision" or "weakly Supervised" or "weakly supervise") AND ("Named Entity" or "Tagging Sequence" or "Sequence Labelling" or "Labelling expression")* no sentido de diminuir os resultados. Dessa forma, a pesquisa retornou 24 resultados na *Web of Science* e 54 na *Scopus*.

A.1.2 Etapa 2: Apresentação e interrelação dos dados

Dada a pequena quantidade de artigos retornados, não foi aplicada seleção temporal como filtro da pesquisa. Conforme pode ser observado nas figuras A.2 e A.3 houve um alto interesse dos pesquisadores nessa área nos últimos anos.

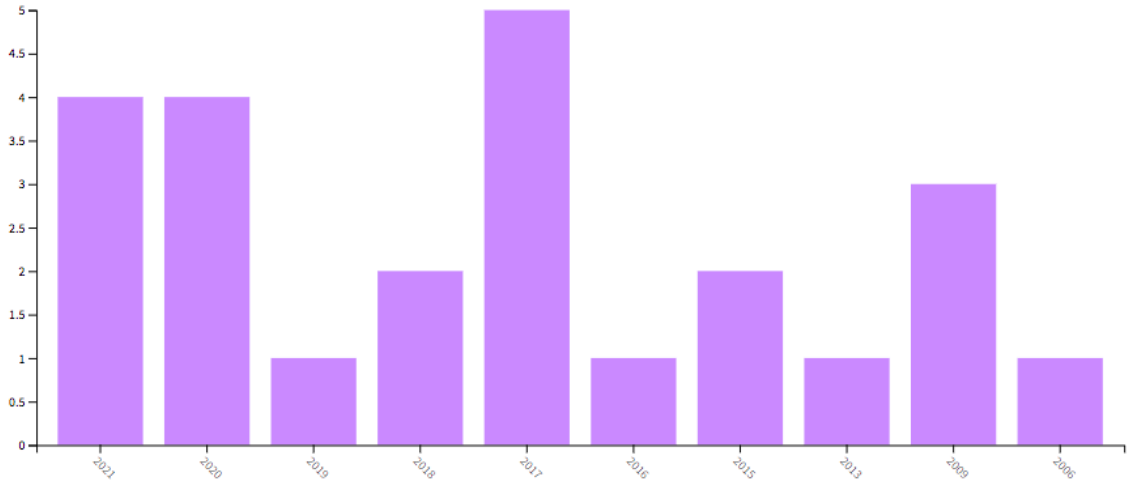


Figura A.2: Publicações por Ano *Web of Science*

As editoras que mais publicaram sobre o tema na *Web of Science* foram: *Assoc Computacional, IEEE, Springer Nature*. Enquanto que na *Scopus* foram: *Lecture Notes In Computer Science, Procedings of The ACM SIGKDD e Ceur Workshop Proceedings3*.

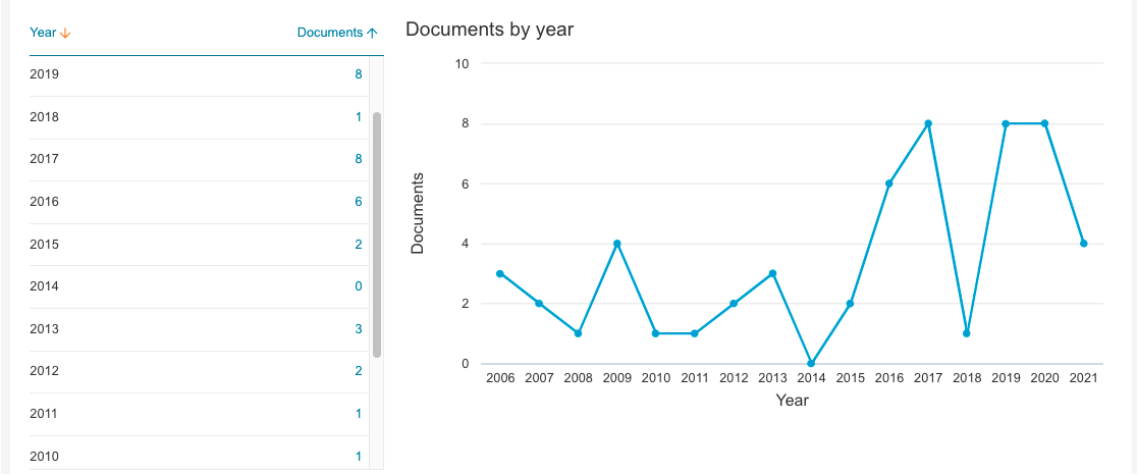


Figura A.3: Publicações por Ano *Scopus*

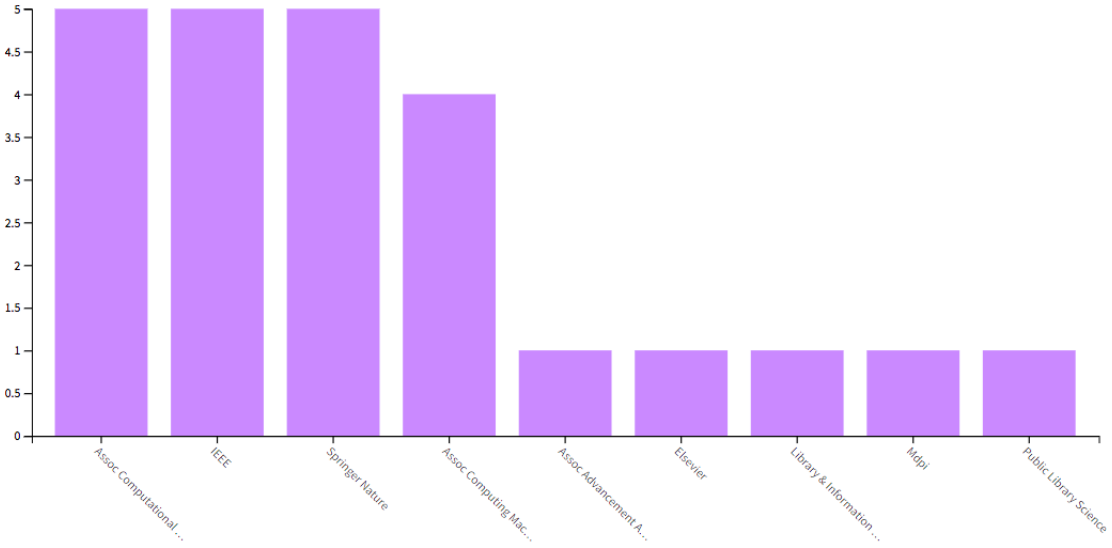


Figura A.4: Editoras que mais publicaram *Web of Science*

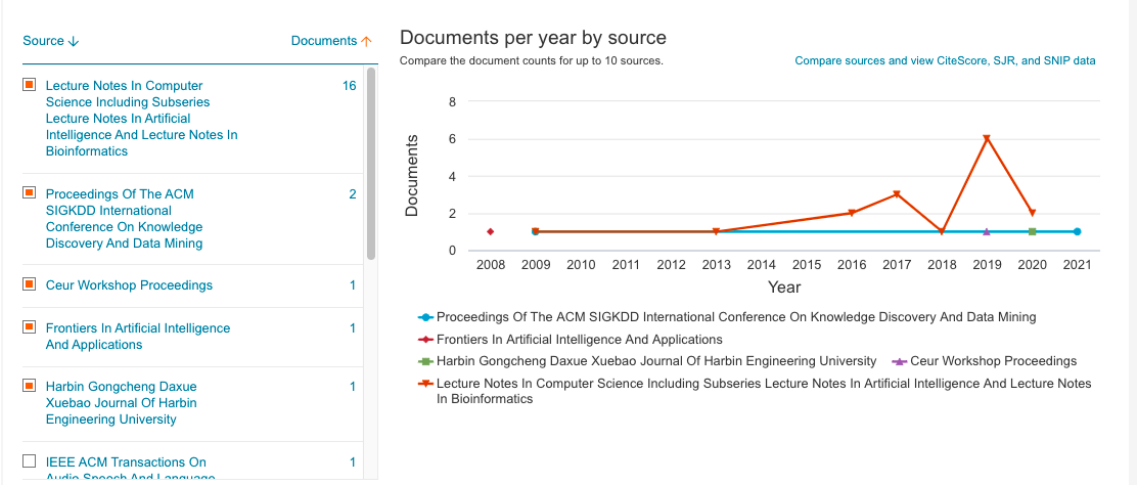


Figura A.5: Editoras que mais publicaram Scopus

Uma forma de estimar o grau de relevância dos periódicos científicos, em determinada área de conhecimento, é através do fator de impacto. Consultando o fator de impacto no *Citation Report do Web Of Science* observou-se que o fator de impacto das publicações é *H-index* 7. Já as publicações no *Scopus* possuem *H-index 15*.

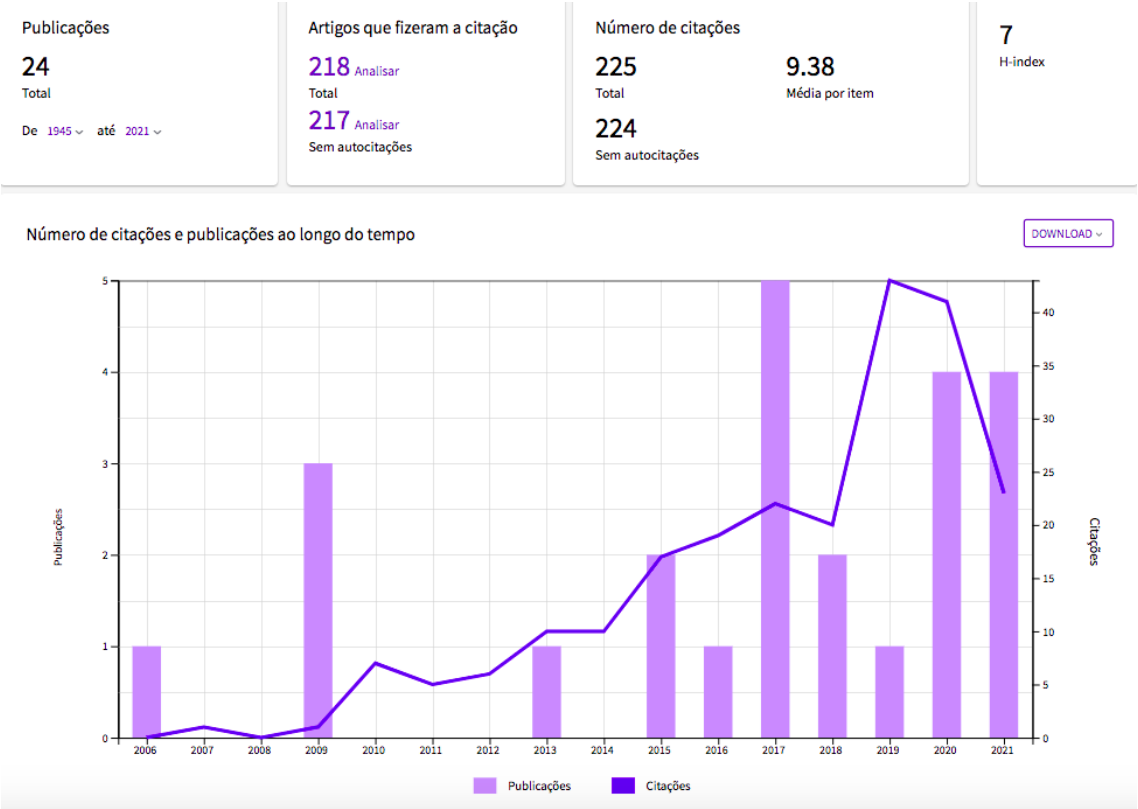


Figura A.6: Gráfico com o número de citações ao longo do tempo *Web of Science*

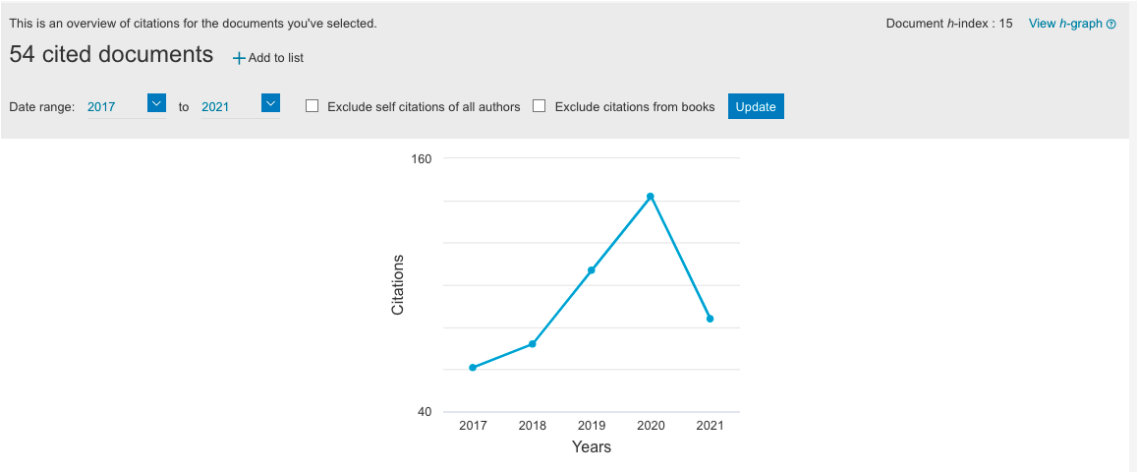
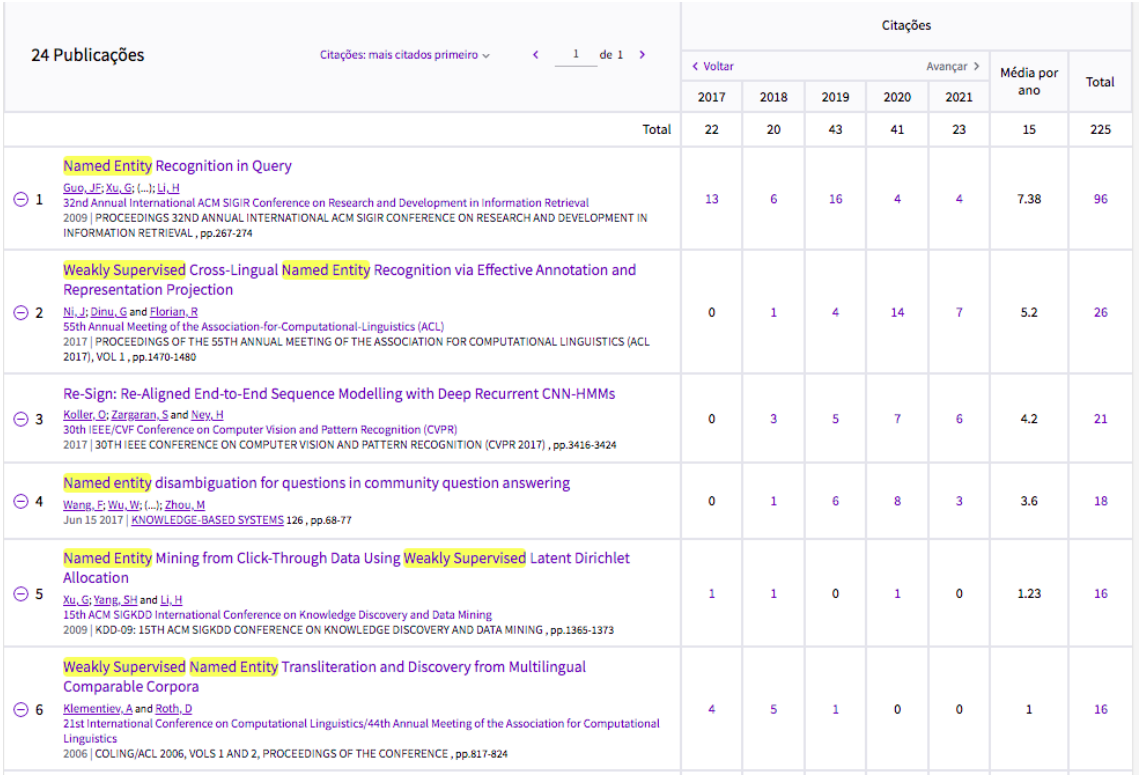


Figura A.7: Gráfico com o número de citações ao longo do tempo Scopus

Nas figuras A.8 e A.9 são apresentados os documentos mais citados na *Web of Science* e no *Scopus*, respectivamente. Em ambas, o artigo mais citado foi o de *Guo, J et al* (2009)



24 Publicações — Citações: mais citados primeiro — 1 de 1

		Citações						
		‹ Voltar				Avançar ›	Média por ano	Total
		2017	2018	2019	2020	2021		
	Total	22	20	43	41	23	15	225
1	Named Entity Recognition in Query Guo, JF; Xu, G; (...); Li, H 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2009 | PROCEEDINGS 32ND ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL , pp.267-274	13	6	16	4	4	7.38	96
2	Weakly Supervised Cross-Lingual Named Entity Recognition via Effective Annotation and Representation Projection Ni, J; Dinu, G and Florian, R 55th Annual Meeting of the Association-for-Computational-Linguistics (ACL) 2017 | PROCEEDINGS OF THE 55TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL 2017), VOL 1 , pp.1470-1480	0	1	4	14	7	5.2	26
3	Re-Sign: Re-Aligned End-to-End Sequence Modelling with Deep Recurrent CNN-HMMs Koller, O; Zargaran, S and Ney, H 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2017 | 30TH IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR 2017) , pp.3416-3424	0	3	5	7	6	4.2	21
4	Named entity disambiguation for questions in community question answering Wang, F; Wu, W; (...); Zhou, M Jun 15 2017 | KNOWLEDGE-BASED SYSTEMS 126 , pp.68-77	0	1	6	8	3	3.6	18
5	Named Entity Mining from Click-Through Data Using Weakly Supervised Latent Dirichlet Allocation Xu, G; Yang, SH and Li, H 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2009 | KDD-09: 15TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING , pp.1365-1373	1	1	0	1	0	1.23	16
6	Weakly Supervised Named Entity Transliteration and Discovery from Multilingual Comparable Corpora Klementiev, A and Roth, D 21st International Conference on Computational Linguistics/44th Annual Meeting of the Association for Computational Linguistics 2006 | COLING/ACL 2006, VOLS 1 AND 2, PROCEEDINGS OF THE CONFERENCE , pp.817-824	4	5	1	0	0	1	16

Figura A.8: Artigos com o número de citações ao longo do tempo *Web of Science*

	Documents	Citations	<2017	2017	2018	2019	2020	2021	Subtotal	>2021	Total
		Total	532	61	72	107	142	84	466	2	1000
1	Named entity recognition in query	2009	136	18	8	23	14	10	73		209
2	Weakly-supervised discovery of named entities using web sear...	2007	88	4	8	3	3		18	1	107
3	Structured relation discovery using generative models	2011	49	5	6	7	8	4	30		79
4	Re-sign: Re-aligned end-to-end sequence modelling with deep ...	2017		2	9	17	34	15	77		77
5	Labeling the languages of words in mixed-language documents ...	2013	35	8	6	10	12	3	39		74
6	Organizing and searching the world wide web of facts - Step ...	2007	70	1		1	1		3		73
7	Weakly supervised named entity transliteration and discovery...	2006	45	5	4	1	1	1	12		57
8	Weakly supervised cross-lingual named entity recognition via...	2017			3	5	24	10	42		42
9	Weakly supervised approaches for ontology population	2006	33		2	1	1	1	5		38
10	Named entity mining from click-through data using weakly sup...	2009	30	2	1		3		6		36
11	Weakly supervised tweet stance classification by relational ...	2016	1	5	9	8	4	2	28		29
12	Docred: A large-scale document-level relation extraction dat...	2020					9	15	24		24
13	Named entity disambiguation for questions in community quest...	2017			2	6	7	6	21		21
14	A weakly-supervised detection of entity central documents in...	2013	14		1	2	2		5		19
15	Data-driven information extraction from Chinese electronic m...	2015		3	5	2	3	1	14	1	15
16	Knowledge-based biomedical word sense disambiguation with ne...	2017			2	8	3	1	14		14
17	A scalable machine-learning approach for semi-structured nam...	2010	9	3	1			1	5		14
18	Weakly supervised approaches for ontology population	2008	11	1	1	1			3		14
19	Web-derived resources for web information retrieval: From co...	2009	10						0		10
20	Weakly supervised sequence tagging from noisy rules	2020					2	6	8		8

Figura A.9: Artigos com o número de citações ao longo do tempo Scopus

Guo et al. (2009) que explora a detecção da entidade nomeada em uma determinada consulta e classifica em classes pré-definidas. Na *Web of Science* o artigo registrou 96 citações e na *Scopus* 209 citações.

Em relação aos países que mais publicaram nas duas bases de dados, como mostrado nas figuras A.10 e A.11, têm-se os Estados Unidos e a China.

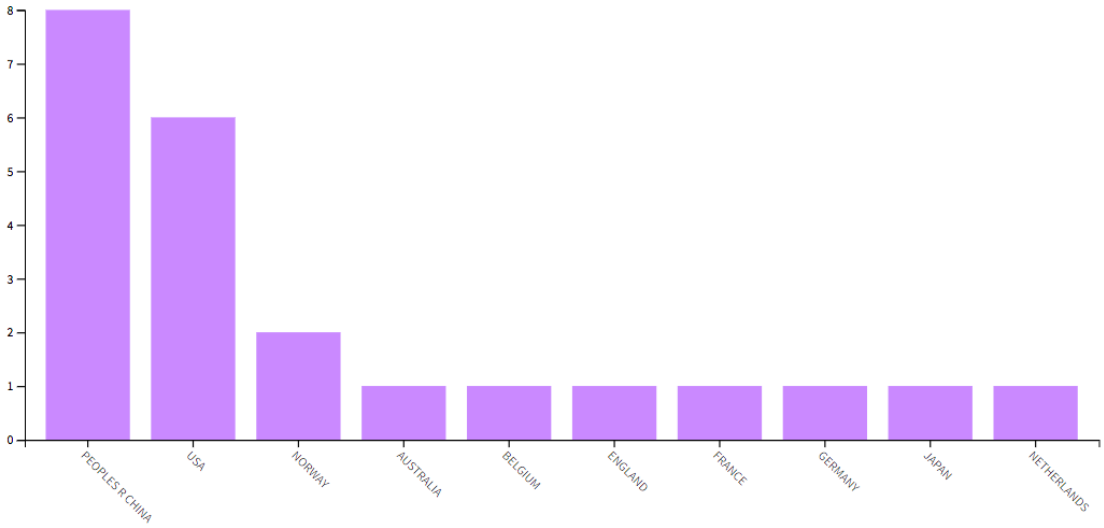


Figura A.10: Paises que mais publicaram *Web of Science*

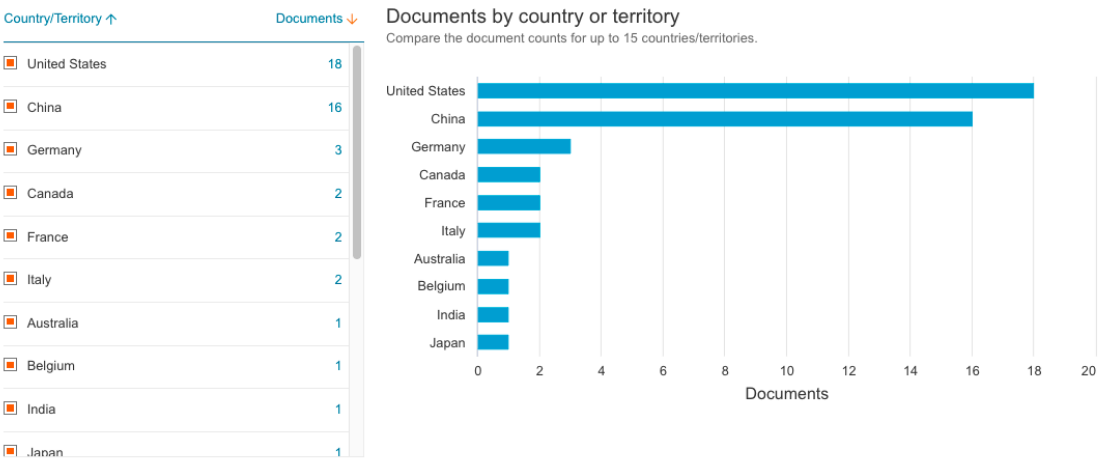


Figura A.11: Paises que mais publicaram Scopus

A.1.3 Etapa 3: Detalhamento, modelo integrador e validação por evidências

Com base nos dados extraídos na *Web of Science* e *Scopus*, foi utilizado o *software VOSviewer* para criação de mapas calor, facilitando a visualização da análise sobre as heurísticas utilizadas na Rotulação de Dados com a abordagem de supervisão fraca. Esses mapas usam cores mais quentes e fontes em negrito para enfatizar conceitos que são frequentemente usados, enquanto palavras que são usadas apenas esporadicamente são mostradas em cores mais frias e fontes menores Zupic (2015).

Na *Web Of Science* foram encontrados apenas 2 grupos de co-autores e que publicaram pelo menos dois documentos sobre assunto. O primeiro grupo composto por: *Hubin, Aliaksandr; Barnes, Jeremy; Lison, Pierre*; publicaram os artigos *"Named Entity Recognition without Labelled Data: A Weak Supervision Approach"* Lison et al. (2020) e *"skweak: Weak Supervision Made Easy for NLP"* Lison et al. (2021).

O segundo grupo composto por: *Li, Hung; Xu, Gu;* publicaram os artigos *"Named Entity Mining from Click-Through Data Using Weakly Supervised Latent Dirichlet Allocation"* Xu et al. (2009) e *"Named Entity Recognition in Query"* Guo et al. (2009).

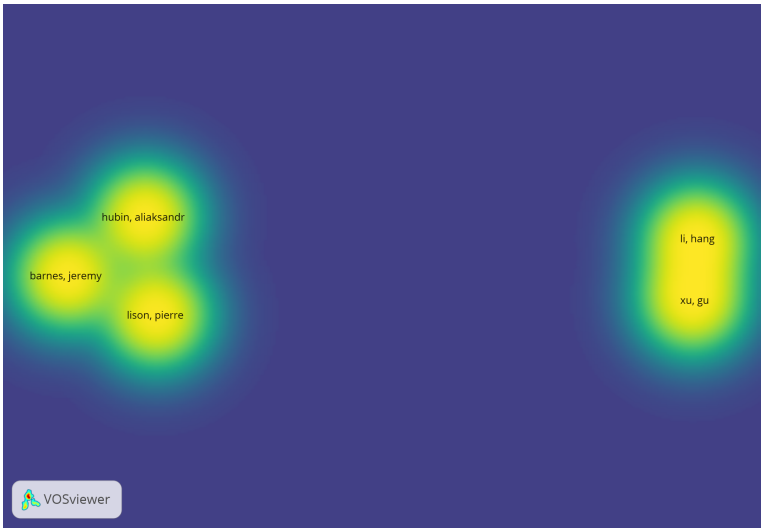


Figura A.12: Publicações Co-Autoria *Web of Science*

Na *Scopus* foram encontrados apenas 7 grupos de autores que publicaram pelo menos dois documentos sobre assunto, porém apenas em 3 grupos foram mapeadas co-autoria. O primeiro grupo composto por: *Li, Z.; Zhang, M.;* publicou o artigo *"Coupled POS tagging on heterogeneous annotations"* Li et al. (2017). Individualmente, *Li, Z.;* publicou *"Named entity disambiguation for questions in community question answering"* Wang et al. (2017) e *Zhang, M.;* publicou *"A supervised named entity recognition method based on pattern matching and semantic verification"* Gao et al. (2020).

O segundo grupo composto por: *Li, H.;Xu, G.;* publicaram os artigos *"Named Entity Mining from Click-Through Data Using Weakly Supervised Latent Dirichlet Allocation"* Xu et al. (2009) e *"Named Entity Recognition in Query"* Guo et al. (2009).

O terceiro grupo composto por: *Wang, D.;Wang, X.;* publicaram *"A weakly supervised Chinese medical named entity recognition method"* Zhao et al. (2020). Individualmente, *Wang, X.;* publicou *"PENNER: Pattern-enhanced Nested Named Entity Recognition in Biomedical Literature"* Wang et al. (2019), e *Wang, D.;* publicou *"Weakly-supervised named entity extraction using word representations"* Deng et al. (2017)

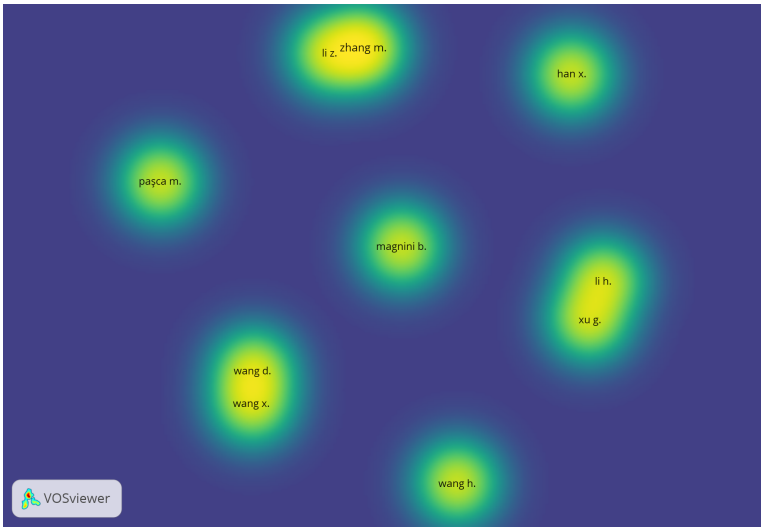


Figura A.13: Publicações Co-Autoria *Scopus*