PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

DANIEL A. GUIMARÃES DE L. REYES

EXTRAÇÃO DE RELAÇÃO ENTRE ENTIDADES NOMEADAS NO CONTEXTO ECONÔMICO-FINANCEIRO

Porto Alegre
2021

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EXTRAÇÃO DE RELAÇÃO ENTRE ENTIDADES NOMEADAS NO CONTEXTO ECONÔMICO-FINANCEIRO

DANIEL A. GUIMARÃES DE L. REYES

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Prof[a]. Isabel Harb Manssour

Porto Alegre
2021

Ficha Catalográfica

D278e De Los Reyes, Daniel Alessandro Guimarães

Extração de relação entre entidades nomeadas no contexto econômico-financeiro / Daniel Alessandro Guimarães De Los Reyes. – 2021.

100.

Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Isabel Harb Manssour.

1. Extração de relação. 2. Extração de relação financeira de entidade nomeada. 3. Extração de relação semântica. 4. Processamento de linguagem natural. I. Manssour, Isabel Harb. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

DANIEL A. GUIMARÃES DE L. REYES

EXTRAÇÃO DE RELAÇÃO ENTRE ENTIDADES NOMEADAS NO CONTEXTO ECONÔMICO-FINANCEIRO

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado(a) em 30 de Agosto de 2021.

BANCA EXAMINADORA:

Profª. Drª. Soraia Raupp Musse (PPGCC/PUCRS)

Profª. Dra. Vládia Célia Monteiro Pinheiro (PPGIA/UNIFOR)

Profª. Isabel Harb Manssour (PPGCC/PUCRS - Orientadora)

DEDICATÓRIA

Dedico este trabalho a minha família.
À minha filha Ana Cecília.
À Jamile, minha companheira.
Meu pai e minha irmã e todos que me acompanharam.

“A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original.”
(Albert Einstein)

AGRADECIMENTOS

À minha filha Ana Cecília, que todos os dias me ensina como ser melhor. Por muitas vezes trazer leveza em meio às preocupações do dia-a-dia. Ainda por ela, que deu força e vontade de continuar, insistir e não desistir. À minha namorada, Jamile, que demonstrou tamanha resiliência, paciência, companheirismo e amor ao longo dessa caminhada. Ao meu pai e minha irmã, que sempre estiveram ao meu lado me apoiando e dando forças. Às minhas orientadoras Profa. Dra. Renata Vieira e Profa. Dra. Isabel Harb Manssour, pela imensurável paciência em me orientar durante todo o processo. Durante o curso tive o prazer de contar com a ajuda de colegas do PPGCC, Allan Barcelos, Anielle Severo, Douglas Trajano entre outros, meu mais sincero obrigado.

EXTRAÇÃO DE RELAÇÃO ENTRE ENTIDADES NOMEADAS NO CONTEXTO ECONÔMICO-FINANCEIRO

RESUMO

Inteligência Competitiva (IC) é uma área relevante de uma corporação e pode apoiar a área estratégica de negócios, auxiliando os responsáveis pela tomada de decisões e como posicionar sua organização no mercado. No domínio financeiro, a identificação das organizações contidas em uma notícia pode se tornar insuficiente, sendo necessário extrair relações (ER) entre as entidades. Assim sendo, o objetivo deste trabalho é propor uma abordagem para a extração de qualquer relação semântica entre Entidades Nomeadas (ENs) no domínio do Mercado Financeiro para a língua portuguesa. Para atingir este objetivo, inicialmente foi feita uma revisão do estado da arte que levou à análise de 76 artigos para identificar as técnicas e conjuntos de dados usados para avaliá-las. Este estudo demonstrou que existem poucas abordagens para a tarefa de ER na língua portuguesa. Portanto, seguindo a metodologia de *Knowledge Discovery in Databases* (KDD) criada por Fayyad, propusemos uma abordagem em cinco etapas, que vai desde a coleta de dados até a avaliação dos resultados. Esta abordagem usa dois modelos baseados em *Bidirectional Transformer Encoding Representations* (BERT) para processar uma frase e suas entidades nomeadas. Primeiro classificamos se um determinado par de entidades tem ou não uma relação semântica e, em seguida, extraímos as partes da frase que representam ou descrevem a relação semântica entre essas entidades nomeadas. A abordagem foi desenvolvida para a língua portuguesa, considerando o domínio financeiro e explorando representações linguísticas profundas sem utilizar outros recursos léxico-semânticos. Os resultados dos experimentos mostram uma precisão de 76,3% usando a métrica de Jaccard, que mede a similaridade entre as relações extraídas pelo modelo extrator, além de alcançar pontuações de 87%, 84,5% e 85,8%, respectivamente para as métricas de *Recall*, Precisão e *F-Measure* quando mensuramos a abordagem completa. Outra contribuição importante é

o corpus construído manualmente com mais de 9.114 tuplas (frase, entidade, entidade) anotadas em *tweets* e notícias disponibilizadas por analistas de IC para apoiar a decisão.

Palavras-Chave: extração de relação, extração de relação financeira de entidade nomeada, extração de relação semântica, processamento de linguagem natural.

EXTRACTION OF RELATION BETWEEN NAMED ENTITIES IN THE ECONOMIC-FINANCIAL CONTEXT

ABSTRACT

Competitive Intelligence (CI) is a relevant area of a corporation and can support the strategic business area, helping those responsible for decision making and how to position your organization in the market. In the financial domain, identifying the organizations contained in a news story can become insufficient, and it is also necessary to extract relations (ER) between entities. Therefore, the main goal of this work is to propose an approach for the extraction of any semantic relation between Named Entities (NEs) in the Financial Market domain for the Portuguese language. To achieve this goal, a state-of-the-art review was initially carried out, which led to the analysis of 76 articles to identify techniques and datasets used to assess them. This study shows that there are readings for the RE task in Portuguese language. Therefore, following the methodology of Knowledge Discovery in Databases (KDD) created by Fayyad, we proposed a five-step approach, which goes from collecting data to evaluating the results. This approach uses two models based on Bidirectional Transformer Encoding Representations (BERT) to process a sentence and its named entities. We first classify whether or not a given pair of entities has a semantic relation and then extract the sentence parts representing or describing the semantic relation between these named entities. The approach was developed for the Portuguese language, considering the financial domain and exploring deep linguistic representations without using other lexical-semantic resources. The results of the experiments show an accuracy of 76.3% using the Jaccard metric, which measures the similarity between the relations extracted by the extractor model, in addition to achieving scores of 87%, 84.5% and 85.8%, respectively for the Recall, Precision and F-Measure metrics when assessing the complete approach. Another important contribution is the manually built corpus with more than 9,114 tuples (phrase, entity, entity) annotated from tweets and news provided by CI analysts to support the decision.

Keywords: relation extraction, financial named-entity relation extraction, semantic relation extraction, Natural Language Processing.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ACE – Automatic Content Extraction Corpus
API – Interface de Programação de Aplicação
BERT – Bidirectional Transformer Encoding Representations
BI-LSTM – Bidirectional Long Short-Term Memory Networks
BI – Business Inteligence
CRF – Conditional Random Fields
EN – Entidade Nomeada
EI – Extração de Informações
ER – Extração de Relações
IC – Inteligência Competitiva
KDD – Knowledge Discovery in Databases
LSTM – Long Short-Term Memory Networks
NYT – New York Times 10 Corpus
OPEN IE – Open Information Extraction
PLN – Processamento de Linguagem Natural
POS – Part of Speech
QP – Questão Principal de Pesquisa
QPS – Questões de Pesquisa Secundárias
REN – Reconhecimento de Entidades Nomeadas
RNA – Redes Neurais Artificiais
RNC – Redes Neurais Convolucionais
RNR – Redes Neurais Recorrentes
SEMEVAL – Semantic Evaluation
SVM – Support Vector Machine

SUMÁRIO

1. INTRODUÇÃO

A tarefa de Extração de Relações (ER) visa identificar e classificar as relações semânticas que ocorrem entre (pares de) entidades reconhecidas em um determinado texto [WSM+20, Tan19, WWMW20, CGC+20]. Extrair relações entre entidades nomeadas do texto é um grande desafio na Extração de Informações (EI), dado o conhecimento de linguagem necessário e a sofisticação das técnicas de processamento de linguagem empregadas. Ao mesmo tempo, esta tarefa pode contribuir para o desenvolvimento de diversas áreas, tais como Sistemas de Perguntas e Respostas, sumarização, entre outras [CPVV14].

Há um interesse crescente por ER, motivado principalmente pelo crescimento exponencial das informações disponíveis na Web, o que pode inviabilizar a tarefa de busca e utilização de tamanha quantidade de dados manualmente. Este contexto torna a ER uma área de pesquisa ainda mais complexa e relevante [EFC+11, WLZL20]. As pesquisas sobre ER para a língua inglesa estão em estágio mais avançado do que para a língua portuguesa. Assim, embora muitos trabalhos possam ser encontrados na literatura sobre ER para inglês [LYL+18, FPPR19, PIW+17, CX20, WSM+20, Tan19], poucos artigos abordam ER para português.

Em relação ao mercado financeiro, domínio abordado neste trabalho, a notícia traz informações sobre setores da economia, políticas industriais, aquisições, parcerias empresariais, entre outros. Automatizar o processo de análise desses dados, na forma de relatórios financeiros, manchetes e anúncios corporativos, pode apoiar a tomada de decisões econômicas pessoais e corporativas [ZZ18]. Desse modo, por exemplo, é possível extrair uma relação de aquisição entre entidades do tipo organização, na qual uma determinada organização (a primeira entidade) foi adquirida (relação) por outra organização (segunda entidade) [Sar08]. Portanto, a pesquisa de ER entre entidades financeiras é a base para a extração automática de informações financeiras que podem ser usadas para auxiliar nas atividades econômicas individuais e na tomada de decisões econômicas nacionais [ZZ18].

As organizações financeiras que precisam monitorar constantemente as movimentações do seu segmento de mercado podem se beneficiar da ER, utilizando-a para automatizar e auxiliar o monitoramento dessas movimentações. Atualmente, grandes empresas contam com um setor de Inteligência Competitiva (IC) responsável por ler atentamente inúmeras notícias sobre organizações para destacar possíveis movimentações do mercado. Estima-se que com um sistema que filtre automaticamente Entidades Nomeadas e as relações entre elas, possa diminuir o esforço e o tempo de trabalho dispendido para essas funções além de, alimentar os sistemas de *Business Inteligence* (BI). Então, também é possível estabelecer uma base de dados com os acontecimentos de mercado, que permita armazenar o conhecimento sobre essas movimentações, de modo que fiquem organizadas.

Entretanto, ao contrário da língua inglesa, que possui um maior número de conjuntos de dados disponíveis para este tipo de pesquisa, a língua portuguesa carece desse tipo de recurso. Na verdade, pelo nosso conhecimento com a pesquisa realizada, não há um grande conjunto de dados preparado para essa tarefa para a língua portuguesa voltada para o domínio do mercado financeiro, o que mostra algumas das dificuldades para o avanço nesta área de pesquisa.

Existem diversas técnicas voltadas para EI, e técnicas de aprendizado profundo têm se destacado recentemente, principalmente devido à sua capacidade de descobrir padrões implícitos nos dados. A literatura tem apresentado algoritmos de aprendizado profundo como *Transformers* [VSP+17], Redes Neurais Recorrentes [LYL+18, WWMW20] e Redes Neurais Convolucionais [YSW20, YCC+20] como boas alternativas, pois foram aplicados de forma eficiente em várias tarefas de processamento de texto sequencial, incluindo a tarefa de ER. Para a língua portuguesa existem alguns trabalhos que abordam a tarefa de ER como os de Cruz e Weitzel [CW18] e de Collovini et al. [CGC+20]. Porém, não temos conhecimento da existência de trabalhos que abordem essa tarefa utilizando técnicas de aprendizado profundo para a língua portuguesa e para o domínio financeiro.

Considerando este contexto, este trabalho visa responder à seguinte pergunta de pesquisa: Como é possível executar a tarefa de Extração de Relações para o domínio do Mercado Financeiro e o idioma Português? Portanto, o objetivo deste trabalho é propor uma abordagem para a extração de qualquer relação semântica entre Entidades Nomeadas (ENs) no domínio do Mercado Financeiro para a língua portuguesa. Para atingir este objetivo foram realizadas as seguintes atividades:

I. Revisão sistemática da literatura contemplando a leitura completa de 76 artigos entre 279 previamente retornados pelas bibliotecas digitais pesquisadas;

II. Elaboração de um processo em formato de *pipeline* com 2 modelos para processar uma frase e suas entidades nomeadas;

III. Criação de um conjunto de dados composto por 9.114 tuplas (frase, entidade, entidade) anotadas manualmente a partir de *tweets* e notícias disponibilizadas por analistas de IC para suporte à decisão;

A abordagem foi estruturada da seguinte maneira: primeiro classificamos se um determinado par de entidades tem ou não uma relação semântica e, em seguida, extraímos as partes da frase que representam ou descrevem esta relação semântica entre essas entidades nomeadas. Usamos o *Bidirectional Transformer Encoding Representations* (BERT), um modelo pré-treinado com a arquitetura de transformadores [VSP+17], de modo que é possível aproveitar o poder do BERT e obter a semântica das frases sem o uso de seleção aprimorada de recursos ou outros recursos externos. *Transformers* BERT foi escolhido pelo fato de já ter sido utilizado na literatura, para a tarefa de ER e com foco

no idioma inglês, apresentando bons resultados [WSM+20, KPGM20, YCC+20, MOM+20, QZZ20, WWMW20]. Outra motivação para a utilização do BERT é a recente disponibilização de sua versão treinada com um Corpus composto inteiramente por sentenças em português [SNL20], e por já ter alcançado bons resultados em tarefas de PLN como Reconhecimento de Entidades Nomeadas [dSN14, SdSK+20]. Portanto, as principais contribuições deste trabalho são:

- Uma abordagem em formato de *pipeline* com 2 modelos para a tarefa de ER para a Língua Portuguesa no contexto financeiro, baseado em BERT. Ambos modelos que compõem a abordagem, o de classificação [DLRBVM21] e o modelo de extração de relações, apresentaram bons resultados em suas respectivas tarefas.
- Um corpus composto por 9.114 tuplas anotadas manualmente, formadas a partir de notícias do mercado financeiro. Disponível no GitHub [1].

O restante deste trabalho está organizado da seguinte forma: O Capítulo 2 contém a fundamentação teórica necessária para o entendimento do método de IA proposto para a tarefa de ER, analisando o contexto do processamento automatizado de informações para a área de IC, e explicando os transformadores de BERT. O Capítulo 3 apresenta uma revisão sistemática da literatura a partir do ano de 2015, referente ao tópico de Extração de Relação entre Entidades Nomeadas. O Capítulo 4 fornece uma descrição detalhada da solução proposta. O Capítulo 5 explica o processo experimental em detalhes, seguido pelo Capítulo 6, que mostra os resultados obtidos e fornece algumas discussões. Por fim, o Capítulo 7 apresenta nossas conclusões, bem como possibilidades para trabalhos futuros.

[1] https://github.com/DanielReeyes/financial-market-corpus

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os principais conceitos relacionados ao trabalho proposto e à Revisão Sistemática da Literatura 3: Processamento de notícias do mercado financeiro como um problema abordado, apresentado na Seção 2.1; alguns algoritmos tradicionais de aprendizado de máquina utilizados para ER nos trabalhos revisados são apresentados na Seção 2.3; Ontologia é definido na Seção 2.4; Uma breve descrição sobre aprendizado profundo e redes neurais profundas (Convolucional e Recorrente) é apresentada na Seção 2.5; BERT, *Transformer* utilizado na abordagem proposta, também é apresentado Subseção 2.5.4; Por fim, os métodos de avaliação utilizados nos experimentos são apresentados na Seção 2.6.

2.1 Inteligência Competitiva e Processamento de Notícias

Hoje, as maiores empresas do segmento financeiro possuem um setor de IC. Por meio dele, informações de diferentes fontes são trabalhadas estrategicamente, permitindo antecipar tendências de mercado e possibilitar a evolução do negócio em relação aos seus concorrentes. Este setor geralmente é formado por um ou mais profissionais dedicados especificamente ao acompanhamento das movimentações dos concorrentes.

IC é um modelo de pesquisa estruturado para fatos não estruturados e análise de dados para apoiar a tomada de decisão da empresa em seu planejamento estratégico. Essencialmente, a IC envolve a coleta legal de informações sobre os concorrentes e o ambiente geral de negócios. O conhecimento obtido com essas informações é então usado para aumentar a competitividade da própria organização [Wei02]. Em tempos de competitividade baseada no conhecimento e na inovação, a IC permite que as empresas exerçam a pró-atividade. As conclusões obtidas neste processo permitem à empresa saber se realmente continua competitiva e se existe sustentabilidade para o seu modelo de negócio. A IC pode trazer algumas vantagens para as empresas que a utilizam, tais como: minimizar surpresas dos concorrentes, identificar oportunidades e ameaças, obter conhecimentos relevantes para a formulação do planejamento estratégico, entender as repercussões de suas ações no mercado, entre outros.

O processo de captura de informações por meio de notícias ainda requer muito esforço manual. Grande parte dessas informações da área financeira aparece na forma de texto livre, contando com o processamento manual desses dados, longe de acompanhar o ritmo de crescimento dos dados, e não podendo que se faça pleno uso desta informação [ZZ18]. Muitas vezes depende de um profissional responsável por ler atentamente inúmeras notícias sobre organizações para destacar possíveis movimentos do mercado, e

esse profissional também pode ficar com esse conhecimento para si mesmo. Espera-se então que, com um sistema que filtra automaticamente as relações entre as entidades do mercado financeiro, o esforço e o tempo despendidos nessas atividades possam ser reduzidos. Outro benefício alcançado é que este mesmo sistema pode alimentar sistemas de BI e estabelecer um banco de dados histórico com eventos de mercado. Assim, o conhecimento sobre os movimentos do mercado pode ser armazenado e organizado.

2.2 Tarefa de Extração de Relações

A tarefa de ER do texto é um dos principais desafios da EI, dado o conhecimento necessário da linguagem e a sofisticação das técnicas de processamento de linguagem empregadas. É o estágio da EI, que visa identificar e classificar relações semânticas que ocorrem entre entidades reconhecidas em um determinado texto [BB07, JM09]. Relações semânticas são relações entre conceitos ou significados envolvendo diferentes unidades linguísticas e componentes [Abr14]. De forma geral, a tarefa de ER não possui um conjunto padrão de relações-alvo, mas um sistema de ER deve ser capaz de reconhecer e extrair relações do tipo “cargo-organização”, “localização de”, "parceria com"entre outras como nos exemplos a seguir:

- A American Express **fez um acordo com a** Amazon para lançar um cartão para pequenas e médias empresas .
- A Havanna **fecha parceria com o** Santander para inaugurar um novo modelo de negócios .
- Rubem Novaes, **presidente do** Banco do Brasil não consegue mostrar serviço.

Para Banko e Etzioni [BE08], os dois principais tipos de ER são domínio fechado e domínio aberto: sistemas ER de domínio fechado consideram apenas um conjunto fechado de relações entre dois argumentos, enquanto sistemas ER de domínio aberto não precisam de uma definição pré-especificada da relação. Um dos desafios da ER de domínio aberto é que, devido à diversidade de relações extraídas, há uma dificuldade em organizá-las seguindo alguns critérios [CMV16a].

Existem alguns aspectos que comumente são analisados para auxiliar na tarefa de ER que podem trazer benefícios e aumentar a eficiência dos métodos empregados, conforme apresentado a seguir [Abr14].

- A ocorrência de palavras que podem expressar uma relação particular em torno ou próximas às Entidades Nomeadas.

- Categorias léxicas providas pela anotação de *Part of Speech* (POS) podem auxiliar a identificar se uma palavra define uma relação ou não.

- Estruturas sintáticas da sentença podem expressar uma relação, tais como sintagmas preposicionais ou sintagmas verbais anotados por um parser.

Estes aspectos podem trazer benefícios aos métodos empregados, mas necessitam de um grande poder computacional. Há diferentes tipos de abordagens para realizar ER, tais como: técnicas de aprendizado supervisionado utilizando corpus anotado; abordagens não supervisionadas com base em padrões de extração genéricos; métodos semi-supervisionados, tais como *bootstrapping*, que necessita apenas de poucos exemplos anotados, e também a abordagem *Open Information Extraction* (Open IE) para extração das relações não definidas previamente.

2.2.1 *Open Information Extraction (Open IE)*

De acordo com Fader et al. [FSE11] e Xavier et al. [XdLS15], a EI tradicional é baseado no treinamento de um extrator com algumas relações previamente definidas. A principal desvantagem da EI tradicional é sua baixa cobertura e seu bom ajuste quando aplicado a um domínio particular. Quando os textos são de domínios diferentes, a intervenção humana pode ser necessária. Para superar esse problema, o *Open Information Extraction (Open IE)* surgiu para extrair fatos sem determinar um conjunto de relações semânticas anteriormente. [GC18].

O *Open IE* é uma tarefa para extrair relações semânticas em textos simples sem determinar previamente essas relações. Os sistemas *Open IE* são geralmente aplicados a soluções na escala da web, como melhorar os sistemas de resposta a perguntas, construções de ontologias, filtragem de documentos e agrupamento. Sena e Claro [SC20] desenvolveram sistemas *Open IE* para o idioma português e suas abordagens foram compostas por regras de transitividade e simetria com o objetivo de inferir novos fatos implícitos em sentenças simples. Enquanto que, Cabral et al. [CSC20] desenvolveram um sistema multilingual composto por regras morfo-sintáticas afim de classificar as relações semânticas.

2.3 Algoritmos Tradicionais de Aprendizado de Máquina

Técnicas ou algoritmos da área de aprendizado de máquina são usualmente empregados em tarefas preditivas como a tarefa de ER. Essas técnicas preditivas geralmente buscam aprender padrões sobre dados históricos, com a finalidade de realizar inferências

sobre eventos futuros, assumindo-se, assim, que contextos passados, utilizados na construção de um modelo preditivo, possuem características semelhantes ao contexto para o qual é realizada a predição [PY10, LBH+15]. Alguns trabalhos revisados durante a revisão sistemática da literatura utilizaram algoritmos tradicionais como Redes Neurais Artificias e *Support Vector Machine* (SVM) que são abordados a seguir.

2.3.1 Redes Neurais Artificiais

Redes neurais artificiais (RNA) são redes computacionais que simulam, simplificadamente, a rede de células nervosas (neurônios) de um sistema nervoso central biológico. Basicamente uma simulação célula por célula, isto é, neurônio por neurônio [Dan13]. Podem também ser caracterizadas como um modelo computacional, com propriedades particulares incluindo a habilidade de adaptar-se e aprender, organizar dados agrupando-os e operações baseadas em processamento paralelo [KKvdSS96]. Os neurônios trabalham como elementos de processamento nestas camadas e são totalmente interligados a partir da entrada até a saída. O número de neurônios na camada de entrada depende do tipo e do número de variáveis de entrada no conjunto de dados. O número de neurônios na camada de saída depende do número de variáveis de resultado de classe que estão associadas com o problema de classificação a ser estudada [Pen14].

Durante o processo de aprendizado, a rede ajusta estes pesos para conseguir classificar corretamente um objeto. É uma técnica que necessita de um longo período de treinamento, ajustes finos dos parâmetros e de difícil interpretação, não sendo possível identificar de forma clara a relação entre a entrada e a saída. Também é difícil para humanos para interpretar o significado simbólico por trás dos pesos aprendidos e das "unidades ocultas"na rede. Porém, a RNA é capaz de tolerar dados com ruídos e é altamente capaz de classificar dados com os quais não teve treinamento, além de possibilitar processamento paralelo [HKP11].

2.3.2 *Support Vector Machine*

Support Vector Machine (SVM) se baseia na teoria de aprendizado estatístico supervisionado e se apresenta muito eficaz para reconhecimento de padrões em geral além de apresentar uma boa precisão em suas classificações [Pen13, HKP11]. SVM oferece alto desempenho e precisão de classificação entre duas classes, uma vez que é capaz de modelar limites complexos de decisão enquanto procura a melhor função de classificação destas classes [HKP11]. Dados pertencentes a duas classes são transformados em pontos, então SVM distribui em um espaço a maior porção possível de pontos. Em seguida o

classificador traça um hiperplano, uma linha que é encontrada pelo SVM com a ajuda de vetores de suporte e margens, para que seja possível separar tais pontos de acordo com sua classificação, deixando-os próximos pela sua similaridade e maximizando a distância entre as duas classes de dados [HKP11].

2.4 Ontologias

Uma ontologia é um modelo de dados que representa um conjunto de conceitos dentro de um domínio e as relações entre estes. Ontologia é definida como uma especificação explícita de uma conceituação compartilhada que representa o conhecimento por meio de conceitos, relacionamentos e indivíduos [Gru95]. Ontologias são utilizadas para realizar inferências sobre os objetos do domínio e podem descrever indivíduos, classes, atributos e relacionamentos.

Em uma ontologia, os indivíduos são os objetos básicos; as classes são coleções ou tipos de objetos; os atributos se referem às propriedades ou características que os indivíduos podem ter e compartilhar; e, por fim, os relacionamentos representam as formas como os indivíduos podem se relacionar com outros indivíduos. A popularidade das ontologias se deve à grande promessa de compartilhamento e entendimento comum de algum domínio de conhecimento que possa ser comunicado entre pessoas e computadores [MA07].

2.5 *Deep Learning*

Os métodos de *deep learning* ou aprendizado profundo têm sido vistos como uma oportunidade de geração de bons resultados no recente desenvolvimento do Processamento de Linguagem Natural (PLN), como a extração de entidades e relações. O aprendizado profundo é formado pela composição de vários níveis de representação, sendo que o número de níveis é um parâmetro livre que pode ser selecionado dependendo das demandas da tarefa em questão [XQP18]. Construir vários níveis de representação ou aprender uma hierarquia de recursos é uma das principais vantagens das técnicas baseadas em aprendizado profundo.

O aprendizado profundo também possui capacidade avançada de extrair recursos que podem transformar as informações semânticas do texto em um vetor de recursos com baixa dimensão e alta densidade, resolvendo efetivamente o problema da escassez de recursos [PYCX18]. Grande parte dos trabalhos desenvolvidos para PLN com métodos de aprendizado profundo envolvem o aprendizado de representações de vetores de palavras por meio de modelos de linguagem [MSC+13] e a composição de vetores de palavras aprendidas para classificação [CWB+11].

A seguir alguns conceitos que costumam ser vistos em diversos trabalhos que envolvem tarefas de PLN serão descritos com a finalidade de fornecer mais insumos para o entendimento de modelos de aprendizado profundo.

2.5.1 Redes Neurais Convolucionais (RNCs)

A RNC foi originalmente inventada para visão computacional e sempre usada para extrair recursos de imagens [LBBH98]. Dentro da área de processamento de imagem, uma das técnicas mais difundidas para solucionar problemas como diminuição de ruídos, detecção de bordas entre outros é a Convolução. A convolução tem raízes na matemática, e pode ser contínua (que é usada em processamento de sinais, por exemplo) ou discreta (que é usada em processamento de imagens).

Para o computador uma imagem corresponde a números, ou matrizes de números, conforme ilustra a Figura 2.1. O processo de Convolução, então, envolve a aplicação de um *kernel*, conhecido também como filtro, sobre uma imagem, gerando uma nova imagem como resultado. O *kernel* também é uma matriz e, na prática, ele "desliza" sobre toda a imagem, modificando-a. Conforme se deslocam, os elementos do *kernel* são multiplicados pelos elementos dos pixels da imagem, que, por fim, são somados e resultam em um elemento da imagem convolucionada. Uma ilustração desse processo pode ser visto na Figura 2.2.

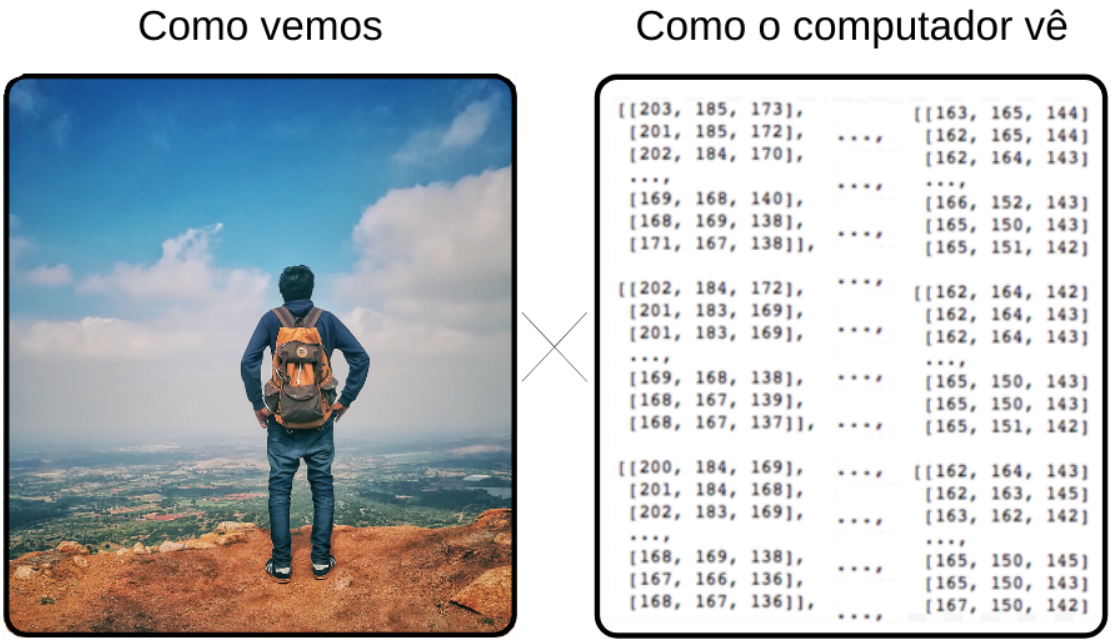

Figura 2.1 – Tradução de uma imagem para linguagem computacional. Fonte: Elaborada pelo autor.

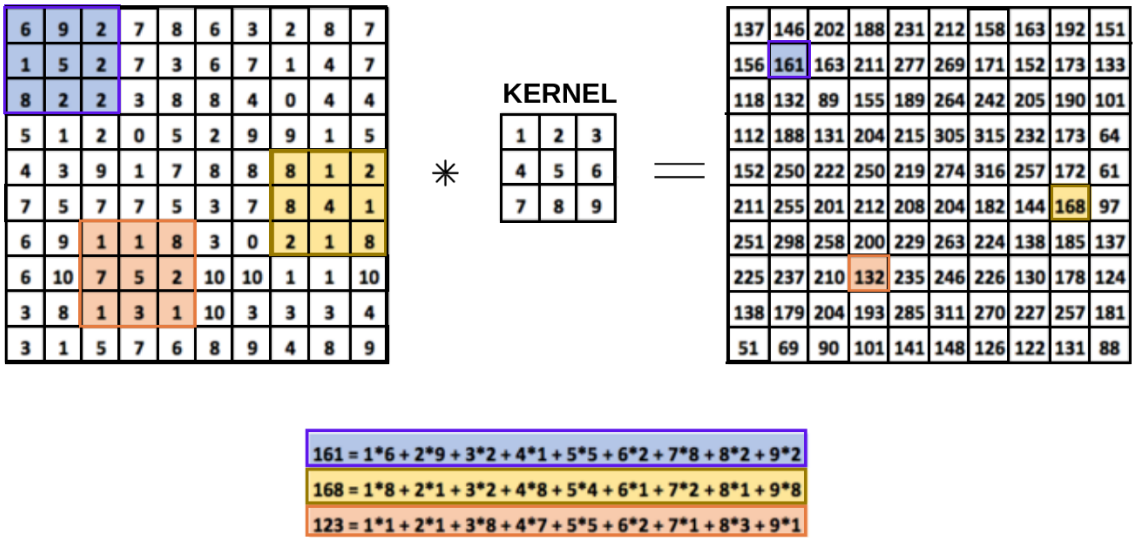

Figura 2.2 – Processo de Convolução: Aplicação de um *kernel* ou filtro em uma imagem. Fonte: Elaborada pelo autor.

Dessa maneira, as RNCs são RNAs que utilizam um tipo especial de camada, apropriadamente chamado de camada convolucional, que faz com que ela obtenha bons resultados em tarefas como classificação de imagens ou textos. Esta camada, possui os *kernels* aprendidos (pesos), que extraem recursos que distinguem imagens diferentes umas das outras. A escolha do tamanho desse hiper-parâmetro tem um enorme impacto na tarefa de classificação de imagens. Tamanhos pequenos de *kernel* são capazes de extrair uma quantidade muito maior de informações contendo recursos altamente locais da entrada, porém demandam maior poder computacional.

Após a aplicação dos filtros, o processo é direcionado para a camada de ativação, onde há o cálculo da ativação da convolução, ou função de ativação. Parte da razão pela qual essas RNCs são capazes de alcançar bons resultados é devido à sua não linearidade. A não linearidade é necessária para produzir limites de decisão não lineares. As funções de ativação aplicam a não linearidade necessária no modelo resultando em um vetor de *features maps*.

Os *features maps* gerados têm tamanhos variados e para que tenham tamanhos fixos uma função é aplicada. Para simplificar e reduzir a dimensionalidade desses vetores, há uma camada de *pooling* que utiliza funções matemáticas (e.g., máximo, mínimo, média) que reduz a dimensionalidade dos *features maps* extraídos. As camadas convolucionais anteriores da rede extraíam os recursos da imagem de entrada que são enviados para a camada *Flat* que irá classificar.

Nos últimos anos, a RNC foi aplicada com sucesso a diferentes tarefas de PLN e demonstrou a eficácia de extrair informações de palavras-chave e semântica de frases [XDFX18], alcançado bons resultados em outras tarefas. Em trabalhos como em [Kim14] e [DSG14], os autores demonstraram que as arquiteturas da RNC podem ser utilizadas

com sucesso na classificação de sentenças. Já em [BAT19] o autor propõe uma técnica utilizando uma RNC para detecção de entidades nomeadas.

Collobert e Weston [CW08] sugerem que podemos tratar efetivamente a matriz de sentenças como uma imagem e realizar convolução nela por meio de filtros lineares. Neste caso, os dados de entrada da RNC são compostos pelas sentenças tokenizadas. Após uma tokenização das sentenças, cada palavra é considerada um *token*, e este será representado por um vetor. Ao final da tokenização cada sentença é transformada em uma matriz S x T onde S se refere a quantidade de palavras e T o tamanho máximo de *tokens*. Como as linhas representam símbolos discretos (ou seja, palavras), é razoável usar filtros com larguras iguais à dimensionalidade dos vetores de palavras [ZW17].

2.5.2 Redes Neurais Recorrentes (RNRs)

RNRs idealizadas por Rumelhart em [RHW86], são a principal ferramenta para lidar com dados sequenciais, que envolvem entradas ou saídas de comprimento variável. Comparado a uma rede de várias camadas, os pesos em uma RNR são compartilhados em diferentes instâncias dos neurônios artificiais, cada um associado a diferentes etapas no tempo. Isso permite aplicar a rede para inserir sequências de diferentes comprimentos, porque os mesmos pesos são reutilizados a cada etapa do tempo [GBC16].

Se tivéssemos que definir uma função diferente para cada comprimento de sequência possível, ou seja, uma rede neural separada, cada uma com um tamanho de entrada diferente, cada uma com seus próprios parâmetros, não teríamos generalização para sequências de tamanho não visto no conjunto de treinamento. Além disso, seria necessário ver muito mais exemplos de treinamento, porque um modelo separado teria que ser treinado para cada tamanho de sequência e precisaria de muito mais parâmetros, proporcionalmente ao tamanho da sequência de entrada.

A Figura 2.3 ilustra a estrutura e o funcionamento das RNRs. As conexões recorrentes permitem que uma memória das entradas persista na rede e, portanto, influencie a sua saída [Gra12]. Essas características fazem com que as RNRs sejam comumente empregadas em diversas áreas de PLN visto que, ao utilizar palavras de uma sentença como entrada, cada uma das computações tem influência sobre a próxima.

Uma rede neural recorrente recebe uma sequência de entrada e realiza o processamento propagando a saída para dentro da rede. Assim, a decisão alcançada na etapa de tempo t-1 afeta a decisão que alcançará um momento mais tarde na etapa de tempo t. As redes recorrentes têm duas fontes de entrada, o presente e o passado recente, que se combinam para determinar como respondem a novos dados.

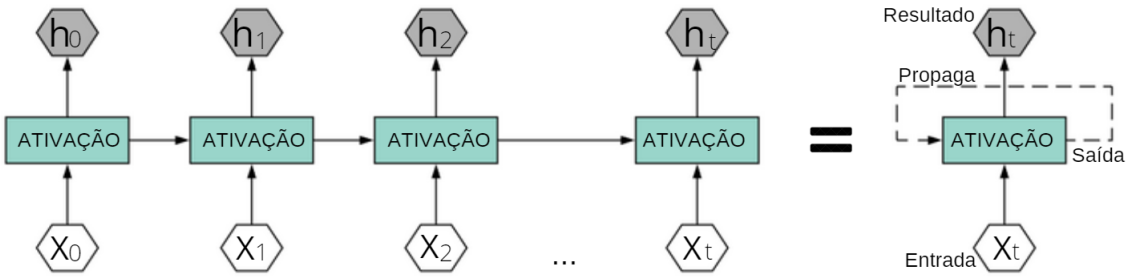

Figura 2.3 – Perspectiva de funcionamento geral de uma Rede Neural Recorrente. Fonte: Elaborado pelo autor, adaptada de Bengio [GBC16]

Contudo, as redes neurais recorrentes simples enfrentam problemas em conservar as memórias de longo prazo. O problema básico é que os gradientes propagados por vários estágios tendem a desaparecer (na maioria das vezes) ou explodir (raramente, mas com muitos danos à otimização). Além disso, quanto menor é o gradiente, mais difícil é para a rede atualizar os pesos e maior é a demora para a convergência. Portanto, o treinamento da rede é comprometido, já que as camadas iniciais não são ajustadas corretamente e realizam a propagação destes valores para as computações subsequentes [KJFF15].

2.5.3 Redes Recorrentes *Long Short-Term Memory Networks* (LSTM)

Em meados dos anos 90, a proposta dos pesquisadores alemães Sepp Hochreiter e Juergen Schmidhuber apresentou uma variação da rede recorrente com as chamadas unidades de *Long Short-Term Memory* (LSTM), como uma solução para o problema do *vanishing gradient*, problema comum em redes neurais recorrentes. Segundo os pesquisadores em [HS97] as redes LSTM conseguem preservar o erro na retro-propagação através do tempo e das camadas. Ao manter um erro mais constante, eles permitem que redes recorrentes continuem aprendendo ao longo de muitos passos de tempo.

Existem inúmeras variantes da LSTM que podem ser encontradas na literatura como [HS97], [Gra12], [GMH13], [GJM13] [SVL14], mas o princípio é sempre ter um *auto-loop* linear através do qual os gradientes podem fluir por longos períodos [GBC16].

Conforme a Figura 2.4, define-se uma unidade LSTM a cada passo de tempo t como uma coleção de vetores contidos num espaço. A coleção é formada por um vetor *input gate*, um *forget gate*, um *output gate*, uma *memory cell* e um estado oculto [TSM15]. Estes portais *"gates"* multiplicativos permitem que as células de memória LSTM armazenem

e acessem informações por longos períodos, mitigando, assim, o problema da dissolução do gradiente [Gra12].

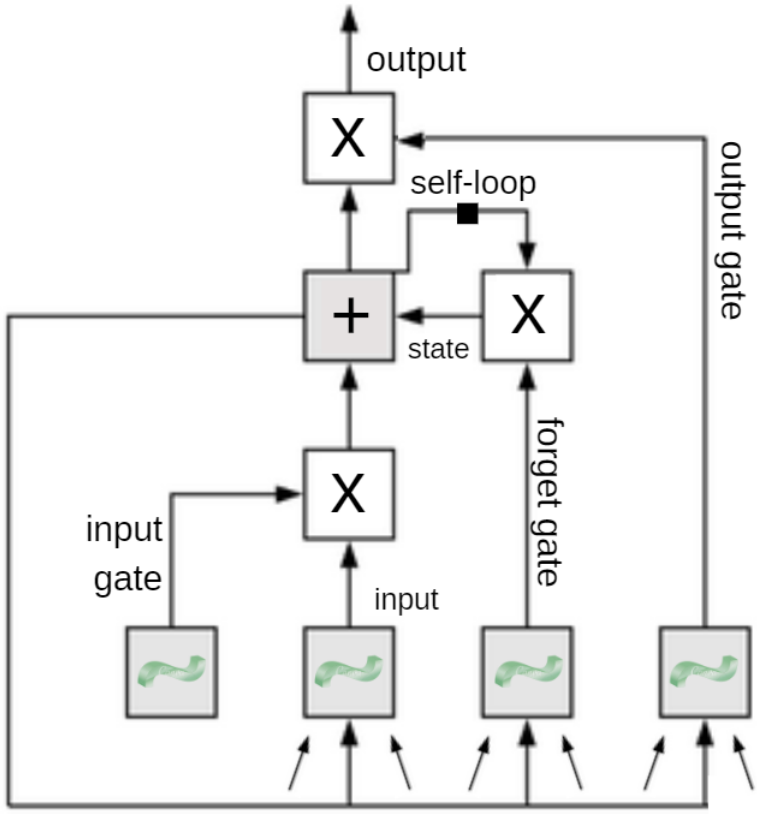

Figura 2.4 – Perspectiva de uma célula de uma Rede Neural Recorrente LSTM. Fonte: Adaptada de [GBC16]

A memória da LSTM contém a estrutura que armazena o estado da célula, e ela percorre a cadeia permitindo remover (esquecer) ou adicionar (lembrar) informações dos estados efetuando algumas interações com os portões, que são compostos de uma ativação Sigmóide e uma multiplicação. Dessa maneira, a função Sigmóide fará o papel de filtrar o que deve ser armazenado no estado da memória.

Seguindo o fluxo de funcionamento descrito por Tai em [TSM15], primeiro decide-se qual informação será armazenada pela unidade. A adição de informações úteis ao estado da célula é feita pelo *input gate*. Para essa tarefa usa-se uma camada Sigmóide, que decide quais valores serão atualizados. Em seguida verifica-se qual informação será descartada da rede, para isso uma camada Sigmóide novamente age sobre os vetores de estado oculto e o vetor de entrada. Duas entradas: x_t (entrada no momento específico) e $h\_t - 1$ (saída de célula anterior) são alimentadas ao *gate* e multiplicadas por matrizes de peso, seguidas pela adição do *bias*. Então, um vetor é criado usando a função de ativação

Tanh que dá saída e contém todos os valores candidatos a novos valores da unidade. Os valores do vetor e os valores regulados são multiplicados para obter as informações úteis.

Enfim, é possível extrair as informações que serão úteis com base no estado anterior no portão de saída. Aplica-se uma camada Sigmóide tomando a decisão de quais partes da unidade serão atualizadas produzindo um vetor de saída. Uma outra camada de tangente hiperbólica recebe o estado *ct* e multiplica pelo vetor *ot* resultando no vetor de saída. Os valores do vetor e os valores regulados são multiplicados para serem enviados como uma saída e entrada para a próxima célula.

Bengio em seu livro [GBC16] indica que as redes LSTM aprendem dependências de longo prazo com mais facilidade do que as arquiteturas recorrentes simples. Primeiro em conjuntos de dados artificiais projetados para testar a capacidade de aprender dependências de longo prazo indicando trabalhos como [BSF94], [HS97] e [HBF+01]. Em seguida, em tarefas desafiadoras de processamento de sequência nas quais foi obtido desempenho de ponta ([Gra12], [GMH13] e [SVL14]).

Seguindo a mesma linha de raciocínio das redes LSTM, também é possível mencionar uma variação da mesma, conhecidas como redes *Bidirectional* LSTM (Bi-LSTM). As redes Bi-LSTM consistem de duas LSTM que funcionam em paralelo. Dada uma sequência de entrada uma LSTM percorre tal sequência em uma direção - para frente – enquanto a outra LSTM percorre da direção inversa - para trás [GJM13].

2.5.4 Modelos Baseados em *Transformers*

Transformers é um modelo de aprendizado profundo introduzido em 2017 por Vaswani et al. [VSP+17] que utiliza o mecanismo de atenção, pesando a influência de diferentes partes dos dados de entrada. Esse modelo possui uma arquitetura codificador-decodificador, na qual o codificador consiste em um conjunto de camadas de codificação que processa a entrada iterativamente uma camada após a outra. Já o decodificador consiste em um conjunto de camadas de decodificação que fazem a mesma coisa com a saída do codificador. Essa arquitetura se baseia inteiramente em mecanismos de atenção para computar as representações de sua entrada e saída.

Os mecanismos de atenção permitem que um modelo olhe diretamente para qualquer ponto ao longo da frase e assim, também permite que o modelo se contextualize com os trechos que se referem a esse ponto. É usado principalmente na área de PLN e suas mais variadas tarefas. O *Transformer* na área de PLN é uma arquitetura que visa resolver tarefas sequencias enquanto lida com dependências de longo alcance com facilidade. A seguir alguns modelos baseados nessa arquitetura são detalhados.

BERT

O BERT é um grande modelo de linguagem pré-treinado proposto pelo Google [DCLT19] em 2018. Desde então, o BERT tem alcançado resultados de ponta em várias tarefas de Processamento de Linguagem Natural [HW20, XZM+19]. O BERT consiste em um codificador *Transformer* multi-camada bidirecional baseado na implementação original descrita em Vaswani et al. [VSP+17]. Cada camada possui duas subcamadas. A primeira contém um mecanismo de auto-atenção com várias cabeças e a segunda é uma rede de alimentação direta totalmente conectada no sentido da posição. As conexões residuais são aplicadas a ambas as subcamadas.

O modelo treina combinando duas tarefas: *Masked LM* e Previsão da Próxima Frase. A primeira tarefa, *Masked LM*, envolve o mascaramento de 15% das palavras de uma determinada frase, cujo valor original deve então ser previsto, considerando o contexto da frase em que aparecem. Destes 15%, 80% é substituído pelo símbolo [MASK], 10% por uma palavra aleatória e 10% é mantido com a palavra original.

No processo de treinamento de BERT, o modelo pega pares de frases como entrada e aprende a prever se a segunda frase do par é a frase subsequente no documento original. Durante o treinamento, 50% das entradas são pares em que a segunda frase é a frase subsequente no documento original, enquanto nos outros 50% uma frase aleatória do corpus é escolhida como a segunda frase. A suposição é que a frase aleatória será desconectada da primeira frase.

Generative Pre-trained Transformer

O *Generative Pre-trained Transformer* (GPT) é um modelo de linguagem criado pela OpenAI capaz de gerar texto escrito. O GPT encontra-se na sua terceira versão e sua arquitetura é uma rede neural baseada em *transformers* como o BERT [RNSS18]. GPT-1 (primeira versão) usou um conjunto de dados com cerca de 7.000 livros não publicados que ajudaram a treinar o modelo de linguagem em dados não vistos. O modelo foi treinado por 100 épocas em lotes de tamanho 64 e comprimento de sequência de 512 e conta com 117 milhões parâmetros no total. Em sua versão mais atual GPT-3 [BMR+20], o modelo é composto por 175 bilhões de parâmetros, o que o torna um modelo mais complexo que suas versões anteriores.

Tentando contornar limitações conhecidas de modelos de aprendizagem supervisionada, como necessidade de grande quantidade de dados anotados para aprender uma tarefa específica que muitas vezes não está facilmente disponível ou dificuldade em generalizar para tarefas diferentes daquelas para as quais foram treinados. O modelo GPT-1 propôs aprender um modelo de linguagem generativa usando dados não rotulados e, em seguida, permite ajustar o modelo, fornecendo exemplos de tarefas posteriores específicas, como análise de sentimento, classificação textual, etc. A aprendizagem não supervisionada

serviu como objetivo de pré-treinamento para modelos supervisionados e ajustados, daí o nome de Pré-treinamento Generativo.

2.6 Avaliação de Métodos de ER

Para que seja possível desenvolver um bom experimento é necessário identificar meios para avaliar se o problema analisado foi devidamente quantificado e se é possível identificar os efeitos da abordagem proposta. Em sistemas supervisionados, a tarefa de ER é expressa como uma tarefa de classificação [MMW10]. Portanto, medidas como Precisão, Abrangência e *F-Measure* podem ser utilizadas para avaliar esses sistemas, uma vez que sistemas supervisionados necessitam de dados de referência para o aprendizado, e esses dados podem ser utilizados para calcular tais medidas.

A avaliação de desempenho de um modelo de classificação é baseada nas contagens de dados de saída correta e incorretamente preditos pelo modelo. Essas contagens são registradas em uma tabela chamada Matriz de Confusão [T⁺06, HKP11]. Em um modelo que se deseja classificar se há uma relação semântica entre entidades nomeadas na sentença, a classificação se torna binária: **contém** ou **não contém**. A Tabela 2.1 ilustra uma matriz de confusão usada para este tipo de problema.

		Resultado Real	
		1 - Contém relação	0 - Não contém relação
Resultado Predito	1 - Contém relação (+)	VP	FP
	0 - Não contém relação (-)	FN	VN

Tabela 2.1 – Matriz de Confusão

A cada classe pode ser atribuído um valor, um positivo e outro negativo [HKP11]. No exemplo da Tabela 2.1, atribuiu-se positivo à classe “contém” e negativo à classe “não contém”. **VP** representa o total de ocorrências de Verdadeiros Positivos, ou seja, tupla que foi corretamente classificada como contendo relação. Enquanto, **VN** corresponde ao total de ocorrências de Verdadeiros Negativos, ou seja, tupla que foi corretamente classificada como não contém relação. **FP** significa o total de ocorrências de Falsos Positivos, ou seja, tuplas que foram incorretamente classificadas como contém relação, pois não contém relação entre as entidades. E por fim, **FN** é o total de ocorrências de Falsos Negativos, ou seja, tuplas que foram classificadas como não contendo uma relação de modo incorreto pois as entidades possuem uma relação semântica entre elas.

Esta matriz possibilita visualizar precisamente a quantidade de predições\classificações corretas e incorretas. A acurácia é resultante da divisão da quantidade de predições corretas pela quantidade de predições feitas. Logo, quanto maior a acurácia, melhor o modelo preditivo. Além dessa métrica é possível analisar a eficiência por meio de outras medidas mais precisas. As medidas comumente utilizadas para avaliações de classificação são Precisão, *Recall* e *F-measure*, definidas da seguinte forma:

$$Recall = \frac{VP}{VP + FN}; \tag{2.1}$$

$$Precisao = \frac{VP}{VP + FP}; \tag{2.2}$$

$$F\text{-}Measure = \frac{2 * Precisao * Recall}{Precisao + Recall}. \tag{2.3}$$

A medida de *Recall* é o percentual de ocorrências classificadas corretamente como positivas dentre todas que realmente são parte da relação. Já a precisão se refere ao percentual de ocorrências classificadas corretamente como positivas dentre todas as que foram classificadas como parte da relação [HKP11]. Por fim a medida de *F-Measure*, que é a média harmônica da precisão e do *Recall*, onde o *F-Measure* atinge seu valor máximo em 1 (perfeita precisão e *Recall*) e o mínimo em 0.

Neste trabalho, também utilizamos a métrica de avaliação Coeficiente de Similaridade de Jaccard ou simplesmente Coeficiente de Jaccard [Jac01]. O coeficiente de Jaccard mede a similaridade entre dois conjuntos e é definido como o tamanho da interseção dividido pelo tamanho da união dos conjuntos de amostra. A equação 2.4 mostra como o coeficiente é calculado e um exemplo de avaliação pode ser visto na Tabela 2.2.

$$jaccard(A, B) = \frac{|A \bigcap B|}{|A \bigcup B|}. \tag{2.4}$$

Sentença	**Relação Verdadeira**	**Relação Extraída**	Pontuação Jaccard
Havanna abrirá cafeteria dentro do **Santander**.	abrirá cafeteria dentro do	abrirá cafeteria dentro	0.75

Tabela 2.2 – Exemplo de avaliação para extrair relações semânticas entre entidades nomeadas. As entidades nomeadas avaliadas estão em negrito.

Para as duas relações da Tabela 2.2, há uma similaridade de Jaccard de $3/(3+1) = 0,75$, que é o tamanho da interseção do conjunto dividido pelo tamanho total do conjunto.

3. REVISÃO DA LITERATURA

Este capítulo apresenta a revisão da literatura. Primeiramente apresentamos uma Revisão Sistemática da Literatura (RSL) sobre o tópico de extrações de relações entre entidades nomeadas nas bases da ACM e IEEE. Visando complementar a pesquisa para encontrar trabalhos que tratem português, foi feita uma busca adicional manual através do Google Scholar especificamente para esse idioma.

3.1 Revisão Sistemática da Literatura

3.1.1 Protocolo da Revisão Sistemática da Literatura

Esta revisão seguiu um fluxo estabelecido por Kitchenham [Kit04] e tem como objetivo apresentar o estado-da-arte a respeito da tarefa de Extração de Relações entre entidades. Para alcançar o objetivo deste estudo, foram definidas cinco questões de pesquisa, a partir de uma pergunta geral sobre o estado da arte:

- Questão Principal de Pesquisa (QP): Qual é o estado-da-arte de Extração de Relações entre Entidades Nomeadas?

A QP pode ser considerada ampla e o intuito não é de contemplar todas as informações para respondê-la. No entanto, foi estabelecido um conjunto de Questões de Pesquisa Secundárias (QPS) para ajudar na identificação de aspectos relevantes para a continuidade do trabalho. O conjunto de questões de pesquisa secundárias é a base do processo de revisão e cada questão foi definida da seguinte maneira:

1. QPS1: Quais as técnicas utilizadas para extração de relações?
2. QPS2: Como é abordada a complexidade das sentenças utilizadas para a extração de relações?
3. QPS3: Quais as principais métricas e metodologias para avaliação dos trabalhos analisados?
4. QPS4: Qual conjunto de dados é utilizado para a avaliação?
5. QPS5: Qual o domínio do corpus utilizado pelos trabalhos analisados?

Identificamos os termos-chaves por meio de uma busca prévia em alguns trabalhos e também utilizando a estrutura de população, intervenção, comparação e resultado

conforme definido por Kitchenham [Kit04], onde a **População**: *“Natural Language Processing"ou “Information Extraction"*; e **Intervenção**: (*“Relation extraction"* ou *“Relation Discovery"*).

As bases escolhidas para buscar os estudos foram: IEEE Explore[1] e ACM Digital Library[2]. Estas foram escolhidas por serem as maiores bases disponíveis nas licenças disponibilizadas pela PUCRS e possuem motores de busca com funções mais fáceis de manusear. Além disso, a maioria dos periódicos e anais das conferências da área é publicada nestas bases. A Tabela 3.1 apresenta as *strings* usadas em cada biblioteca digital e a quantidade de trabalhos retornados. A partir dos trabalhos retornados, foi feita a extração dos dados, a seleção e sumarização dos resultados.

Tabela 3.1 – *Strings* de pesquisa utilizadas nas bases de busca.

Base de Dados	**Quantidade de Trabalhos Resultantes**	**String de Pesquisa**
IEEE Digital Library	205	(*(“Relation extraction” OR “Relation Discovery”) (AND “Abstract":entity))*
ACM Digital Library	74	[*[Abstract: “relation extraction"] OR [Abstract: “relation discovery"]*] *AND [Abstract: “entity"] AND [Publication Date: (01/01/2015 TO 04/30/2021)]*

Para garantir que os trabalhos revisados sejam recentes, sobre tópicos relacionados à pesquisa, que apresentem informações confiáveis, e para que sejam de fácil acesso ao leitor, foram definidos os critérios de inclusão e exclusão apresentados na Tabela 3.2.

A seleção dos estudos ocorreu em duas etapas. Na primeira etapa, títulos e palavras-chaves foram lidos, seguido de uma leitura dos resumos de cada artigo para tentar identificar ocorrências fora do escopo da revisão aplicando os filtros de inclusão e exclusão. Na segunda etapa, com os demais trabalhos remanescentes, a partir de uma leitura completa de cada estudo eles também foram filtrados seguindo os critérios de inclusão e exclusão. O resultado dessa etapa de filtragem pode ser visto na Figura 3.1, na qual é possível identificar a atuação de cada filtro em cada base de dados.

Ao fim desse estágio foram identificados 106 trabalhos para em seguida ser iniciada a segunda etapa, que contempla a leitura completa para que dessa maneira possa-se responder as perguntas de pesquisa. Durante a etapa de leitura dos trabalhos remanescentes, foram excluídos mais 30 estudos de acordo com os filtros estabelecidos.

[1] Https://ieeexplore.ieee.org/Xplore/home.jsp
[2] Http://dl.acm.org

Tabela 3.2 – Critérios de inclusão e exclusão de trabalhos.

	Inclusão	**Exclusão**
I	Estudos que abordem Extração de Relação de Entidades	Trabalhos cujas palavras chaves e resumo não estejam focados em Processamento de Linguagem Natural ou áreas correlatas
II	Estudos que abordem técnicas para Extração de Relação	Trabalhos que o termo extração de relações entre entidades não esteja relacionado a Processamento de Linguagem Natural ou Extração de Informação
III	Estudos publicados a partir de 2015	Qualquer estudo que não esteja em um formato de artigo completo (por exemplo, resumos de conferências, livro, capítulos OU resumos estendidos)
IV	Estudos escritos em inglês ou português	Formatos diferentes de PDF
V	Estudos que tratem soluções para o idioma inglês, português ou espanhol	Estudos que não tratem de soluções para idiomas diferentes de inglês, português ou espanhol
VI		Trabalhos secundários como surveys
VII		Trabalhos duplicados

Após a filtragem dos trabalhos há a coleta de dados onde é registrado um conjunto de informações sobre esses estudos. Neste conjunto estão incluídas informações sobre as técnicas utilizadas por cada autor e as medidas de avaliação utilizadas para atestar a sua eficácia, assim como informações sobre os conjuntos de dados, idioma, domínio e se eles são disponíveis para consulta.

A Tabela 3.3 lista todos os trabalhos selecionados após a aplicação do protocolo de pesquisa. Depois que os trabalhos foram selecionados, o próximo passo foi organizar e sintetizar os dados para obter informações úteis. Analisando os dados da Tabela observa-se que houve um aumento no número de trabalhos desta área a partir de 2019, sendo que o domínio geral é o que mais aparece. Também chama a atenção que entre os 76 trabalhos apenas um é para língua portuguesa e o corpus mais usado é o NYT. As respostas para as perguntas de pesquisa encontradas durante a leitura dos trabalhos apresentados na Tabela 3.3 são apresentadas nas seções a seguir.

3.1.2 Quais as técnicas utilizadas para extração de relações?

Os métodos de extração de relações podem ser divididos em métodos baseados em regras e métodos baseados em aprendizado de máquina, de acordo com seu princípio de realização. Entre eles, o método de extração de relações baseado em aprendizado de máquina é o mais usado, visto que ele não é restrito ao domínio e, portanto, pode ser transferido para diferentes áreas e em relação aos sistemas baseados em regras, seu custo de mão-de-obra pode ser menor. O método baseado no aprendizado de máquina é dividido

Citação	Ano	Domínio	Técnica	Idioma	Corpus Público?
[CUKR15]	2015	Geral	SVM	Inglês	Não
[ZZX+15]	2015	Geral	Palavras Gatilho	Inglês	Não
[SKL+15]	2015	Saúde	Regras	Inglês	Não
[HZSBHM15]	2015	Saúde	Ontologia	Inglês	SemEval-2013
[ZLW+16]	2016	Geral	Rank	Inglês	Wikipedia
[LJHZ16]	2016	Geral	RNR + BERT	Inglês	Wikipedia
[DJ16]	2016	Geral	RNC	Inglês	Não
[LCW16]	2016	Geral	RNC	Inglês	NYT
[FGQ+17]	2017	Geral	RNC	Inglês	NYT
[JLHZ17]	2017	Geral	RNC	Inglês	NYT
[NP17]	2017	Saúde	SVM	Inglês	Não
[YGJ+17]	2017	Geral	RNC	Inglês	NYT
[ZYS+17]	2017	Saúde	RNC	Inglês	Não
[NTW17]	2017	Geral	CRF	Inglês	Não
[TQG17]	2017	Geral	Semi-Supervisionada - Kernel	Inglês	NYT + Wikipedia
[RWH+17]	2017	Geral	Supervisionada à distancia - Similaridade	Inglês	Não
[CK17]	2017	Geral	Regras	Inglês	Não
[TGQ17]	2017	Geral	Semi-Supervisionada - Bayes	Inglês	NYT + Wikipedia
[SG17]	2017	Geral	RNC	Inglês	ACE2005 + SemEval 2010
[WZCL18]	2018	Geral	RNR - BiLSTM	Inglês	NYT
[SJC+18]	2018	Geral	RNC	Inglês	NYT
[LRW+18]	2018	Geral	RNR - LSTM	Inglês	Não
[LEF18]	2018	Geral	Ontologia e Regras	Inglês	Não
[CW18]	2018	Geral	Sistema ReVerb	Português	Não
[LYL+18]	2018	Saúde	RNR - Bi-LSTM	Inglês	Não
[GYJ+18]	2018	Geral	RNR - Bi-LSTM	Inglês	Não
[SSJ+18]	2018	Geral	RNR - Bi-LSTM	Inglês	SemEval2010
[AJ18]	2018	Geral	Probabilidade	Inglês	Não
[ZZ18]	2018	Financeiro	RNR - Bi-GRU	Inglês	Não
[LHCW19]	2019	Geral	RNC	Inglês	NYT
[Tan19]	2019	Geral	RNR - BiGRU	Inglês	Não
[PLL+19]	2019	Geral	RNR - BiLSTM	Inglês	NYT
[LQP+19]	2019	Geral	RNC	Inglês	SemlEval2010
[PPM+19]	2019	TI	Rede Neural - Feed Forward	Inglês	Não
[ZSZ19]	2019	Geral	RNC + RNR - BiGRU	Inglês	NYT
[YH19]	2019	Geral	RNR - Bi-LSTM	Inglês	SemEval-2010
[FPPR19]	2019	Saúde	RNR - Bi-LSTM	Inglês	MADE 2018
[HZS19]	2019	Geral	RNC	Inglês	NYT
[WLR+19]	2019	Geral	Supervisionada à distancia - RNC	Inglês	NYT
[RFJ19]	2019	Saúde	CNN + RNR - Bi-LSTM	Inglês	SemEval 2013
[CLZZ19]	2019	TI	RNR - Bi-LSTM	Inglês	Não
[PHT+19]	2019	Geral	RNC	Inglês	NYT
[NLW+19]	2019	Geral	RNR - Bi-GRU	Inglês	NYT
[YDZ+19]	2019	Geral	RNR - Bi-GRU	Inglês	NYT
[PST+19]	2019	Geral	RNC + RNR - Bi-GRU	Inglês	NYT
[SHL19]	2019	Geral	RNN - Bi-LSTM	Inglês	SemEval 2010
[CG19]	2019	Saúde	RNN - Bi-LSTM	Inglês	Não
[ZLL+19]	2019	Geral	RNN - Bi-LSTM	Inglês	NYT
[GGH19]	2019	Geral	RNN - Bi-LSTM	Inglês	SemEval 2010
[WXG19]	2019	Geral	RNC + RNR - Bi-LSTM	Inglês	SemEval2010
[ZLWX19]	2019	Geral	RNC + RNR - Bi-GRU	Inglês	NYT
[YDHX20]	2020	Geografia	RNR - BiLSTM	Inglês	Não
[LYC+20]	2020	Saúde	RNR - BiLSTM	Inglês	Não
[AT20]	2020	Saúde	Random Forest	Inglês	MADE 2018
[CX20]	2020	Geral	RNC + RNR - BiLSTM	Inglês	SemEval-2010
[WSM+20]	2020	Geral	RNR + BERT	Inglês	SemEval-2010
[YZZ20]	2020	Geral	RNC + RNR - BiLSTM	Inglês	SemEval-2010
[WXD20]	2020	Geral	RNC + RNR - BiLSTM	Inglês	NYT
[WWMW20]	2020	Geral	RNC + RNR - BiLSTM	Inglês	NYT
[KPGM20]	2020	Geral	Transformers - BERT	Inglês	SemEval-2010
[YSW20]	2020	Geral	RNC + RNR - BiLSTM	Inglês	SemEval-2010
[YCC+20]	2020	Geral	RNC + BERT	Inglês	NYT
[YJTC20]	2020	Geral	RNC	Inglês	NYT
[MOM+20]	2020	Geral	RNC + BERT	Inglês	NYT
[YJW+20]	2020	Geral	Regressão Linear	Inglês	NYT
[QZZ20]	2020	Educação	Transformers + BiGru	Chinês	Não
[WLYL20]	2020	Geral	RNC + BERT	Inglês	ACE05
[ZZ20]	2020	Geral	RNC	Inglês	SemlEval2010
[WC20]	2020	Geral	RNC	Inglês	NYT
[HZL20]	2020	Geral	Supervisionada à distância - RNR - BiLSTM	Inglês	Não
[PAA20]	2020	Geral	Manual	Inglês	Wikipedia
[CWY+20]	2020	Geral	RNC	Inglês	SemEval 2010
[DRR20]	2020	Saúde	RNR - Bi-LSTM	Inglês	SemEval 2010
[TKK21]	2021	Saúde	RNC + RNR - BiLSTM	Inglês	TAC DDI 2018
[LLS+21]	2021	Geral	RNC	Inglês	ACE05
[CT21]	2021	Geral	RNR + BERT	Inglês	NYT

Tabela 3.3 – Lista dos trabalhos selecionados.

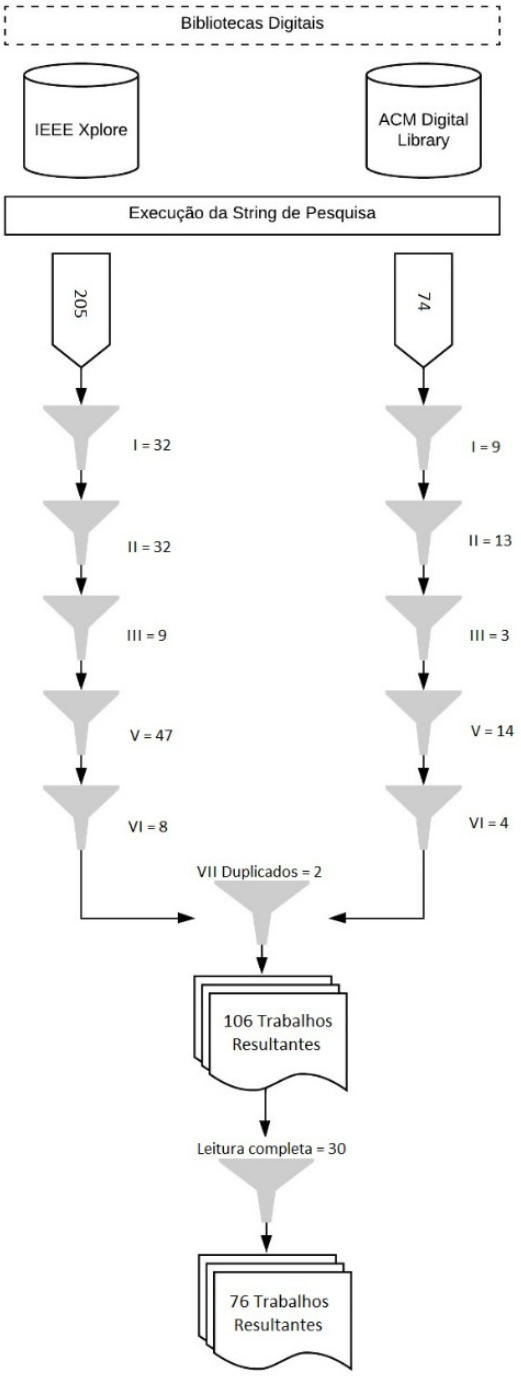

Figura 3.1 – Processo de filtragem dos trabalhos detalhado por filtro de exclusão.

em método de aprendizado supervisionado, método de aprendizado semi-supervisionado, método de aprendizado não supervisionado e aprendizado profundo.

Sistemas baseados em Regras: Dependem de um esforço humano muito grande para que se faça um rastreamento de padrões de regras de escrita. Tais sistemas, também conhecidos como sistemas baseados no conhecimento, consistem em definir heurísticas na forma de expressões regulares ou padrões linguísticos. Sistema baseados em regras também fazem o uso de dicionários ou léxicos que contêm, comumente, a ocorrência de termos ou palavras gatilhos. Song et al. [SKL+15], Romadhony [RWP15], Chaniago [CK17],

Rinaldo Lima [LEF18], Caio e Leila [CW18], são exemplos de abordagens baseadas em regras. Estes autores utilizaram recursos como POS, árvores de dependência, tipos de entidades, divisão de sentenças, tokenização e lematização.

Chaniago [CK17] utiliza estes recursos em conjunto com palavras chaves como regras, por exemplo, palavras chaves como *irmã*, *irmãos*, *irmão*, *gêmeos*. Ou palavras chaves como *parte de* indicando obras, membro, companheiro. Diferentemente, Romadhony [RWP15] também acaba usando árvores de dependência em conjunto com simplificação de sentenças para extrair da melhor maneira possível as relações entre entidades. Caio e Leila [CW18] adaptam um sistema nativo do idioma inglês para o idioma português, ReVerb, baseado em regras para tentar extrair relações entre entidades nomeadas. O ReVerb utiliza informações semânticas para produzir um conjunto de triplas no formato (arg1, rel, arg2), que devem satisfazer às restrições sintáticas e léxicas. Também utiliza um dicionário de relações que está no idioma Inglês, assim a estratégia adotada nesta pesquisa foi fazer a tradução deste dicionário através da API Google Translate. Por fim, utiliza um classificador com regressão logística com objetivo de filtrar triplas espúrias ou, então, que não agregue informação atribuindo um valor de confiança para cada extração.

Song et al. [SKL+15] extrai entidades baseadas em dicionários utilizando um componente externo, o Stanford CoreNLP, [3]. Com o auxílio de regras baseadas em árvore de dependência identifica o verbo da frase, e em seguida, usa uma lista de bio-verbos classificada para determinar a relação entre entidades focadas especificamente no domínio biomédico. Rinaldo Lima [LEF18] também utiliza o Stanford CoreNLP para capturar as informações da sentença. O autor apresenta OntoILPER um sistema conjunto para extrair entidades e relações usando ontologia e programação lógica indutiva. Ele induz regras de extração que incluem exemplos de entidades e instâncias de relação. Conta com a ideia de que o caminho mais curto entre duas entidades pode representar a relação entre elas. Além do Stanford CoreNLP, o autor utiliza componentes externos como o *Rule Learning* para aprendizado de regras capaz de aprender conceitos complexos não determinados.

Classificação supervisionada: é um processo usual pelo qual se implementa uma solução para uma tarefa preditiva, e consiste em duas fases: treinamento e teste [HKP11]. Na fase de treinamento, emprega-se um algoritmo para aprender um modelo classificador sobre um conjunto de dados. Este algoritmo então, tem por objetivo estabelecer uma associação entre tuplas de entrada com seus respectivos rótulos de classe. Para que isso ocorra, deve-se apresentar na fase de treinamento um corpus previamente anotado, para que ele consiga reconhecer padrões entre as tuplas e os rótulos. Na segunda fase, o modelo/classificador estimado é usado para classificar tuplas cujo rótulo de classe é desconhecido, tendo como resultado a inferência de um rótulo para cada nova tupla de entrada.

Técnicas de classificação supervisionada são comumente utilizadas em casos onde se possui uma quantidade maior de dados e estes possuem uma classificação. Zit-

[3] https://stanfordnlp.github.io/CoreNLP/

nik [ŽB15] utiliza algoritmo de *Conditional Random Fields* (CRF) alinhado com uma ontologia, um algoritmo probabilístico, juntamente com recursos como tokenização, POS, *lemmas* e árvore de dependência. Também utilizando CRF, Nguyen et al. [NTW17] propõem uma abordagem conjunta para desambiguação de entidades e extração de relações. Neste trabalho o autor considera a unificidade de cada entidade para inferir uma relação mais correta. Collovini [CMV16b] também utiliza CRF para extrair descritores de relação, utilizando alguns dos mesmos recursos.

Outro algoritmo base utilizado é a árvore de decisão, aplicado por Lossio [LVHM+16], o qual tem por objetivo identificar se uma entidade *e1* possui relação com uma entidade *e2*. Ilseyar e Elena [AT20] utilizam *Random Forest*, uma derivação de árvores de decisão, para extrair relações com base em recursos adicionais divididos em quatro categorias: (I) com base na distância, (II) baseado em palavras, (III) incorporação e (IV) baseada no conhecimento. Apesar de alcançar bons resultados, o modelo se torna muito custoso, visto que depende de muitas informações adicionais que devem ser geradas no momento do experimento.

Kai-Wei et al. [CUKR15], Pankaj et al. [NP17], Gupta [GM17] fazem uso do algoritmo SVM em seus trabalhos. Gupta propõe um modelo de classificação baseado em recursos, que usa recursos léxicos, de sintaxe, de incorporação de palavras e baseados em evidências médicas para classificar a relação do texto médico. Pankaj et al. propõem um método para indicar se duas entidades possuem ou não relação semântica numa mesma frase. Neste estudo as entidades são partes do cérebro e o autor também utiliza como recurso, POS, árvore de dependência, a frequência das palavras e a criação de um dicionário. Já Kai-Wei procura demonstrar que há possibilidades de melhorar o tempo de treinamento e inferência utilizando teoremas da inferência para evitar chamadas desnecessárias do modelo principalmente na fase de treinamento do modelo.

Estes trabalhos anteriores têm em comum o uso de muitos recursos que necessitam de pré-processamento feito por outros softwares de PLN. Dessa maneira é possível que caso um recurso seja extraído de forma incorreta, este erro seja propagado ao longo do processo. Com o advento dos modelos de representações (*Word Embbedings*) e com o surgimento de algoritmos de aprendizado de máquina mais robustos, como os que se baseiam em aprendizado profundo (*Deep Learning*), foi possível generalizar estes recursos e trabalhar de forma mais assertiva na criação de modelos para ER. Algoritmos como Redes Neurais Convolucionais e Redes Neurais Recorrentes, descritas nas Subseções 2.5.1 e 2.5.2 respectivamente ganharam destaque nos últimos anos, há um aumento na utilização dos mesmos visto que eles conseguem trabalhar bem com dados sequenciais, o que é o caso da tarefa de ER.

Aprendizado profundo: Liang et al. [LCW16] cria um método baseado em RNC com o uso dos recursos POS e árvore de dependência para auxiliar durante a fase de classificação da rede. Woo Do [DJ16] sugere arquitetura de rede neural baseada na RNC

para a classificação da relação temporal na sentença. Utiliza recursos léxicos como marcar as entidades e os tempos verbais. Contudo o autor indica que os resultados ruins podem ser pela ausência de mais informações para o modelo como recursos de POS, árvore de dependência entre outros. Já Shi [SG17] propõe utilizar RNC incorporando a posição de entidades na sentença dividindo-a em partes e treinando uma rede para cada parte. Também utilizando RNC, Zhang et al. [ZLWZ17] propõem incorporar posição de palavras e entidades juntamente com uma abordagem com combinações de N-gramas para extração de relações. Em [CWY+20], os autores criaram uma estrutura multicanal, na qual uma frase é dividida em partes. Cada parte é processada por um canal independente com camadas de RNC empilhadas. Ele permite que cada palavra aprenda representações diferentes na mesma frase.

Tran et al. [TKK21] Utiliza RNC com um mecanismo de atenção e recursos como árvores de dependência para focar em determinadas palavras e categorias de palavras. O mecanismo de atenção seria mais seletivo por possivelmente atribuir um valor mais alto para negações e adjetivos e um valor mais baixo para artigos. A Rede Convolucional é usada para representação de caractere e também para identificar o contexto da relação. Cheng e Xiong [CX20] também propõem um método de extração de relação que apresenta um mecanismo de atenção e dois canais de rede neural. A extração de recursos é realizada usando uma camada BiLSTM, descrita na Subseção 2.5.3, enquanto a rede neural convolucional introduz o mecanismo de atenção para detectar onde o modelo deve se atentar, e finalmente, a última camada contendo uma função SoftMax é usada para a classificação do tipo de relação. Lige et al. [YZZ20] propõem um modelo de Rede Neural Convolucional em conjunto com rede BiLSTM e as chamadas cápsulas, no qual a cápsula é um vetor, que pode conter qualquer valor. No entanto, o algoritmo dinâmico é similar a uma rede totalmente conectada, dessa maneira, se torna extremamente custoso quando se trata de grandes corpus.

Liu et al. [LLS+21] têm como objetivo propor uma estrutura geral baseada em RNC de grafos, para extrair relações entre entidades múltiplas em texto não estruturado. A estrutura explora correlações e propagação de informações entre palavras e relações em uma RNC de grafos para compreender características fundamentais para a classificação final. Ren [RFJ19] também utiliza RNC em grafos tratando a tarefa como um problema de previsão de relação tripla e constrói o grafo de entidade enumerando todos os possíveis candidatos. Utiliza *Word Embbeding Glove 300* para converter a sentença e passar a mesma para a camada BiLSTM. Em seguida o modelo verifica todas relações candidatas para um par de entidades via camada de extensão e uma vez que uma relação válida é confirmada pelo SoftMax, a tripla é gerada. Wang et al. [WLYL20] usa BERT, descrito na seção 2.5.4, para obter representação de cada palavra da sentença. A partir dessa representação o autor consegue detectar as entidades na frase. Em seguida utiliza essas informações em uma camada BiLSTM, para obter a representação dessas entidades. Por fim, utiliza essas

informações em uma Rede Neural Convolucional de Grafos sobre árvores de dependência junto com mecanismo de atenção para gerar o vetor de contexto e capturar informações a serem processadas por uma última camada totalmente conectada para extrair as relações entre as entidades.

Outro algoritmo de *Deep Learning* são as redes neurais recorrentes. Mais recentemente, para melhorar a performance desses modelos gerados por algoritmos de aprendizado de máquina notou-se a utilização de mecanismos de atenção. Estes mecanismos auxiliam na etapa de filtrar automaticamente as informações que melhor indicam a descrição que mais perfeitamente se enquadra com a relação da sentença. Desta maneira é possível que mesmo em uma sentença muito grande e pelo seu tamanho ser considerada complexa, o modelo consiga capturar as informações do contexto de cada *token* na sentença e consiga concentrar mais os pesos de influencia nestes *tokens*.

Wu [WH19] apresenta um método para extrair as relações para o idioma inglês apenas utilizando um modelo de linguagem pré-treinado BERT e a informação do tipo de entidades. Enquanto que, Jin Liu et al. [LRW+18] propõem um modelo baseado em LSTM que utiliza informações das sentenças. O modelo proposto utilizou recursos que compreendem entidade, posição da entidade e POS Tag. Yu Wang et al. [WSM+20] anexa caracteres especiais para marcar as entidades a serem testadas e também utiliza o modelo pré-treinado BERT para efetuar a tarefa de ER.

Qingqing Li et al. [LYL+18] utiliza redes neurais recorrentes Bi-LSTM para seu modelo multitarefa, e apresenta uma versão com mecanismo de atenção que melhora consideravelmente os resultados em todos os datasets testados. Florez [FPPR19] utiliza tipos e palavras das entidades que estão sendo consideradas para uma relação, informações sobre entidades como número de entidades, além de informações sobre distâncias como número de palavras e frases entre o par de entidades. A entrada da camada bidirecional LSTM é uma sequência de incorporação de palavras para cada relação, com todas as palavras entre as entidades candidatas (incluídas), fornecidas por uma camada de incorporação pré-treinada. Já Yi [YH19] propõe juntar um modelo pré-treinado de representação de palavras e uma rede neural Bi-GRU que é uma versão da Bi-LSTM com menor custo computacional. Treinam modelo com base no modelo pré-treinado BERT, em vez de treinar do zero, o que pode acelerar a cobertura.

Pengda Qin [QXG17] propõe um método utilizando RNN Bi-GRU com um mecanismo de atenção que é capaz de se concentrar automaticamente em palavras valiosas, baseado em pares de entidades, que tira proveito suficiente das informações do par de entidades. A GRU bidirecional é utilizada para capturar informações valiosas no nível da palavra. Luoa [LYC+20] também propõe um modelo de Rede Neural Recorrente Bi-LSTM, em conjunto com uma camada de *Word Embedding* ELMO e uma camada de atenção para focar em contextos importantes, além da última camada conter um CRF. Combinam

o modelo proposto com a elaboração de regras de marcação de *tokens*. Estas regras se mostraram de grande valia para entidades sobrepostas.

Pandey et al. [PIW+17] usam o mecanismo de atenção no topo da RNN para gerar pesos de atenção a cada passo. No topo do BiLSTM com mecanismo de atenção, usam uma camada CRF para decodificar em conjunto rótulos de toda a frase para detectar a presença de ADE. Tao Gan [GGH19] utiliza uma rede LSTM de atenção a entidades em nível de subsequência para se concentrar mais nas informações importantes de contexto entre duas entidades para a relação expressa em frases brutas. Deepa et al. [DRR20] extraem relações entre duas entidades nomeadas que se estendem por várias frases usando redes LSTM. O sistema proposto considera termos intermediários entre entidades para ER. Utiliza uma função de ativação Softmax para identificar qual a relação com melhor probabilidade. Por fim, também é possível apresentar o trabalho de Zhou [ZZ18] em que o autor implementa um modelo baseado em RNN Bi-Gru com mecanismo de atenção para focar nas premissas mais importantes das sentenças para o mercado financeiro.

Classificação semi-supervisionada: o método de extração de relações semi-supervisionadas pode não apenas reduzir a escala dos dados de marcação artificial, mas também obter um melhor efeito de extração, sendo amplamente utilizado. É indicado quando temos certeza de alguns padrões iniciais e então podemos utilizar métodos como *bootstrapping* e procurar novos padrões nos dados de entrada. Dessa forma é possível gerar um contexto mais generalizado. O algoritmo de *bootstrapping* é um dos algoritmos de aprendizado semi-supervisionados amplamente utilizados; ele precisa apenas de um pequeno número de padrões de sementes para iterar e aprender muitos outros padrões. Apesar de ser bom quando temos poucos padrões confiáveis, a indicação de um padrão incorreto para inicialização pode levar a geração de modelos com pouca eficiência.

O método semi-supervisionado de aprendizado de máquina pode reduzir efetivamente a dependência da participação humana e da anotação de corpus e pode ser estendido à tarefa de extração relacional de texto em larga escala. No entanto, o método *bootstrapping* tem o problema de desvio semântico no processo iterativo, o que afeta a precisão dos resultados da extração.

Uma abordagem semi-supervisionada foi proposta por Liting Tai [TQG17] que utiliza uma função *kernel* adaptada para considerar o peso para o vetor de característica de cada palavra. Função essa baseada em árvore de dependência, para medir a similaridade entre os padrões que entram no *dataset*. O autor também utilizou recursos de análises de palavras como POS e também análise de dependência para pegar o menor caminho de dependência para cada sentença. Liting Tai também propõe um outro método que busca contornar a limitação do algoritmo de *bootstrapping* de desvio semântico, utilizando o conceito de palavra-gatilho. A palavra-gatilho pode ser usada como a âncora semântica do texto e pode restringir efetivamente as informações semânticas do texto.

Moreira et al. [MOM+20] propõem reaproveitar o modelo RESIDE baseado em Rede Neural Convolucional que usa outras informações como tipo de entidade nomeada a ser testada. O reaproveitamento se dá na utilização de uma BERT para representação vetorial das sentenças ao invés do codificador usado no projeto inicial. Assim como Moreira, Yu et al. [YJW+20] abordam extrair a tarefa de ER a distância e sua técnica consiste em combinar palavras de contexto global com tipos de entidades nomeadas mais específicas para poder utilizar regressão linear, método SM2R para fazer o calculo de pontuação candidata.

Zhu [ZSZ19] apresenta uma nova arquitetura de rede neural com um mecanismo de atenção em seu artigo. Usa uma arquitetura GRU bidirecional para codificar as definições de relação como representações contextuais das relações. Em seguida, utiliza o mecanismo de atenção de mesclagem para fazer uso completo do estados ocultos obtidos pelo GRU. Esse mecanismo também ajuda o modelo a aproveitar o contexto das entidades, além de introduzir pesos semânticos, calculados pelo comprimento do caminho mais curto entre entidades.

Classificação não-supervisionada: No caso de classificação não supervisionada, a ideia é agrupar os textos de acordo com suas características, sem nenhum tipo de aprendizado prévio com dados pré-classificados e ainda sem lista prévia de relações semânticas. Dessa maneira acaba possibilitando a extração de relações em um maior número de textos, desde que possua uma estrutura adequada para processar essa grande quantidade e variabilidade de textos. Quan et al. [QWR14] apresentam um método não supervisionado baseado no agrupamento de padrões e análise de sentenças para lidar com a extração de relações biomédicas. O algoritmo de agrupamento de padrões é baseado no método *Polynomial Kernel*, que identifica palavras de interação de dados não rotulados.

3.1.3 Como é abordada a complexidade de sentenças utilizadas para a extração de relações?

Durante os trabalhos revisados, poucos foram os que abordaram a complexidade de sentenças de forma clara. Como referência entre os trabalhos revisados, Ade Romadhony [RWP15] optou por tratar todas as sentenças do seu corpus simplificando-as utilizando regras de dependência. O trabalho propõe simplificar a sentença antes de aplicar a extração de relação. A primeira parte é a etapa de processamento da entidade. A segunda parte é o processamento de cláusulas. O objetivo da primeira etapa é obter uma menção mais curta à entidade e eliminar a sentença que não possui relação de par de entidades.

Isaiah et al. [MSO17] abordou a complexidade das sentenças em seu método de ER para sistemas de perguntas e respostas, de forma a identificar a cardinalidade de relações presentes em uma sentença. Nesse caso uma pergunta é considerada simples

quando a questão expõe apenas uma sentença. Mas neste estudo a ER é apenas um meio para atingir o objetivo de melhorar os resultados de perguntas e respostas. Aditya [PPM+19] escolhe empiricamente um limite de 35 palavras, que é usado como um limite de tamanho da janela para analisar as entidades nomeadas e suas relações. Jin Liu et al. [LRW+18] usa recursos de quantidade de palavras entre outros para descrever relações complexas e melhorar o desempenho do modelo de extração de relação. Nos seus experimentos, classificam o corpus em três conjuntos compostos de 0 a 20 palavras, 20 a 35 palavras e frases com mais de 35 palavras.

Uma forma de lidar indiretamente com um tipo de complexidade de sentença foi observado em outros trabalhos, como de Xu et al. [XDFX18] que usou as redes LSTM como forma de tratar a distância entre entidades. Dessa maneira a LSTM acaba por aprender o contexto da sentença memorizando a sequência de palavras (*tokens*) que se relacionam ao par de entidades e à relação a ser extraída. Shen et al. [SSJ+18] também propõem uma abordagem com uso de redes LSTM para modelar a relação de entidade com o contexto e retém as partes relevantes dos contextos para determinar a relação semântica com a entidade. Florez [FPPR19] acaba utilizando como recurso as palavras presentes entre as entidades em uma camada LSTM de sua rede.

Zhou [ZZ18] constatou que os dados do campo financeiro possuem relações implícitas na sentença e que os textos geralmente são longos. Dessa maneira a mesma semântica pode ser expressa de várias formas. Além disso, como a frase é longa, há muitas informações redundantes. Então o autor acaba abordando tais características com o uso de redes Bi-GRU que são uma versão de rede com características da LSTM, mas com custo menor, além de adicionar mecanismos de atenção a nível de palavra e a nível de sentença. Tao Gan [GGH19] utiliza uma rede LSTM de atenção a entidades em nível de subsequência para se concentrar mais nas informações importantes de contexto entre duas entidades para a relação expressa em frases brutas.

A partir da revisão sistemática, observamos que apenas 3 trabalhos [RWP15, PPM+19, LRW+18] consideraram a complexidade das sentenças de modo explícito. Outros trabalhos relatados envolvem recursos para tratamento implícito da complexidade de sentença, ou seja, ao invés de tratar a complexidade ou mensurá-la diretamente, são empregados recursos para considerar o tamanho de sentença, como as redes LSTM e suas derivações, ou para focar em palavras ou contextos mais relevantes para a ER como os mecanismos de atenção.

3.1.4 Quais as principais métricas e metodologias para avaliação dos trabalhos propostos?

Desde os primeiros estudos, trabalhos que utilizam ER avaliam sua eficiência usando medidas de recuperação de informações. A tripla de medidas precisão, *Recall* e *F-measure* são as medidas mais populares aplicadas a essa área quando se trata da classificação do tipo de relação entre as entidades nomeadas. Essas métricas foram mais difundidas para avaliação de tarefas de ER a partir de sua introdução em 1998 na *Message Understanding Conference* [ZCL17]. Os estudos ainda indicam que a métrica de precisão ainda é mais utilizada do que as outras duas, estando presente em grande parte dos trabalhos revisados. Chen et al. [CG19] chega a citar as três medidas como um padrão para mensurar a assertividade de seu método.

Ainda é possível averiguar que uma quarta métrica é utilizada em alguns trabalhos [TQG17, RWH+17]: a acurácia. Porém, esta medida apesar de indicar a quantidade de relações extraídas corretamente, não é amplamente utilizada pois não apresenta tantos detalhes quando se deseja verificar qual tipo de relação é mais suscetível a falhas ou acertos.

3.1.5 Qual conjunto de dados é utilizado para a avaliação?

A maioria dos estudos apresentam soluções de ER para textos em inglês, e dessa maneira também é possível identificar uma quantidade maior de conjuntos de dados com textos em inglês. Há também uma concentração de estudos que utilizaram amplamente 3 conjuntos de dados em inglês, New York Times 10 (NYT), *Automatic Content Extraction* (ACE) e *Semantic Evaluation* (SemEval), este último tendo suas derivações. Muitos trabalhos [JLHZ17, WZCL18, SJC+18, ZSZ19, WC20, CT21] voltados à extração de relações à distância para o idioma inglês acabam por utilizar o conjunto de dados do NYT. Wu, Zeqiu, et al. também utilizaram o *dataset* NYT para uma abordagem de ER para seu sistema de perguntas e respostas, enquanto Haihong [HZS19] o utilizou para sua abordagem baseada em supervisão a distância com Rede Neural Convolucional por Partes.

O conjunto de dados SemEval também é amplamente utilizado, como em [WXG19, GGH19, DRR20, CX20, YH19, CWY+20]. Entre os trabalhos que utilizam o mesmo conjunto de dados é mais fácil comparar seus resultados, visto que utilizam os mesmos recursos de sentença, diferenciando-se apenas da abordagem proposta. O que se pode notar é que há pouca diversidade de conjuntos de dados públicos. Estes conjuntos são normalmente criados e anotados para convenções como SemEval ou ACE.

A falta desse tipo de recurso força os pesquisadores a desenvolver seus próprios corpus de pesquisa. Isto acaba ocorrendo geralmente quando se trata de domínios muito específicos como ocorre na área da saúde. Assim, é necessário primeiro criar um conjunto com as sentenças e anotá-las quando a classificação for supervisionada para poder prosseguir na tarefa de ER. Tratando-se do domínio da saúde, é possível citar os exemplos de trabalhos de Lossio et al. [LVHM+16], Song et al. [SKL+15] e Pankaj [NP17], onde os mesmos criaram os seus corpus a partir de resumos de artigos da base PubMed. Também para área da saúde, autores como Florez [FPPR19] e Alimovaa [AT20], utilizaram corpus de conferências como *MADE Challenge*. No contexto financeiro, Zhou [ZZ18] apresenta sua abordagem com um *dataset* com aproximadamente 3 mil peças coletadas manualmente de sites de notícias. Wu [WLZL20] aborda em seu trabalho a criação de um corpus para o idioma chinês para extração de relações entre entidades nomeadas, a partir de notícias.

Voltado ao idioma português, há poucos conjuntos de dados disponíveis, como Coleção Dourada do Segundo HAREM, que é utilizado por Chaves [Cha08], Cardoso [Car08] e Collovini et al. [CMV16b]. Porém, vários outros trabalhos utilizam conjuntos de dados de uma variedade de domínios com diferentes recursos, dificultando a sua comparação. A falta de conjuntos de dados públicos também dificulta a comparação justa dos trabalhos relacionados, além de demandar mais tempo e esforço ao pesquisador.

3.1.6 Qual o domínio do conjunto de dados utilizado pelo trabalho proposto?

Ao identificar o domínio do conjunto de dados é possível que se possa fazer uma inferência sobre o domínio que a tarefa de ER pode ser utilizada e o quão generalista o método abordado pode ser. Dessa maneira foi considerado importante responder esta questão de pesquisa. Durante o processo da RSL, foram encontrados muitos conjuntos de dados que se repetiram como visto na subseção 3.1.5 como NYT, SemEval e ACE. Mesmo utilizando alguns desses conjuntos de dados, foram identificados alguns agrupamentos de domínios que serão descritos.

Geral: Foram declarados de âmbito geral os corpus utilizados em trabalhos diversos sem apontar um domínio e que são compostos geralmente por notícias como o NYT utilizado em trabalhos como [HZS19, WXG19, LCW16, PHT+19]. Há também corpus compostos de artigos da Wikipedia como os utilizados em [NTW17, TQG17, AJ18, PAA20]. Apesar de serem da mesma fonte de dados, Wikipedia, geralmente esses corpus possuem características próprias, pois são compostos por artigos que se referem a um ou mais contextos e que são escolhidos aleatoriamente.

Financeiro: Apesar de possuir grande importância o domínio financeiro foi pouco explorado pelos trabalhos revisados. Foi encontrada apenas uma abordagem de ER especificamente voltada ao mercado financeiro. Zhou [ZZ18] criou um corpus coletando 3000

informações manualmente dos principais sites de notícia. Assim, o utiliza para reconhecimento da entidade e a extração de relações como aprendizado e treinamento como um todo. Este estudo em específico se faz importante para trabalhos futuros, mesmo sendo um idioma diferente. A partir do mesmo, é possível elucidar algumas dúvidas como qual o tamanho de conjunto de dados podemos utilizar, qual técnica pode ser utilizada e como podemos medir a sua eficiência.

Saúde: Foi o domínio que apresentou uma maior diversidade entre os conjuntos de dados devido a sua especificidade de assunto de cada trabalho estudado. Lossio [LVHM+16] montou um conjunto de dados bem específico sobre obesidade e câncer para extrair relações entre entidades e, organizar e fornecer informações aos consumidores sobre qualidade de saúde. Florez [FPPR19] utilizou *MADE dataset*, um conjunto de dados pronto de outros trabalhos para abordar a ER entre entidades médicas e assim identificar a relação de reação adversa a medicamentos a partir de notas médicas. Assim como Florez, Pandey [PIW+17] abordou a reação adversa a medicamentos com um conjunto de dados ADE que possui amostras retiradas do MEDLINE e mais um corpus da ADE que possui resumos retirados do PubMed. Ren et al. [RFJ19] utilizaram o DDI corpus da SemEval criados para tópicos de saúde para testar suas abordagens e extrair relações sobre o uso de medicamentos.

Tecnologia: Chen [CLZZ19] criou um corpus para ER entre entidades de erros em código (*bugs*). O autor selecionou aleatoriamente 500 relatórios de erros do projeto Mozilla[4] e extraiu o título, a descrição e os comentários de cada relatório de *bug* para construir o corpus. Nayak [NKD20] também criou um corpus próprio constituído de documentos de software contendo a descrição de todo o sistema de software que consiste em diferentes módulos, funções, estados, sub-estados. Também utilizou declaração de requisitos e relatórios de testes. Com essas informações o autor propõe uma metodologia para criar um grande banco de conhecimento a partir de documentos de engenharia de software que podem ser usados para geração automatizada de casos de teste e declarações de requisitos de idioma natural. Por fim, Aditya [PPM+19] criou o seu próprio corpus para o seu estudo, a partir de textos de cibersegurança coletando informações de boletins, mídias sociais e conjuntos de dados sobre vulnerabilidades.

3.1.7 Discussão dos Trabalhos Analisados

Através das perguntas de pesquisa e suas análises, foi possível verificar que as técnicas para abordar o tema de ER podem ser divididas em 2 tipos, regras e aprendizado de máquina. Ainda é possível dividir técnicas de aprendizado de máquina em mais 4 categorias, como aprendizado supervisionado, semi-supervisionado, não supervisionado

[4] https://bugzilla.mozilla.org/

e aprendizado profundo, o que é determinado principalmente pelo tipo de conjunto de dados disponível para treinamento de um modelo de ER. A Tabela 3.3 ajuda a identificar que entre as técnicas utilizadas predominaram algoritmos de Redes Neurais Profundas e suas variações.

Também por meio dos trabalhos revisados, é possível indicar que a complexidade de sentenças ainda é considerada apenas implicitamente. Os trabalhos empregam determinadas técnicas para poder considerar apenas o tamanho das sentenças, assim como o contexto delas. Apenas 3 trabalhos [RWP15, PPM+19, LRW+18] consideraram a complexidade das sentenças de modo explícito, tentando de alguma forma tratá-la antes da tarefa de ER.

Outra pergunta respondida diz respeito à disponibilidade de dados para elaboração de métodos de ER. Em muitos trabalhos vistos, foi identificada a utilização de um mesmo conjunto de dados, geralmente disponibilizado em eventos de PLN. A Tabela 3.4 sumariza os conjuntos de dados mais utilizados entre os artigos analisados. Os conjuntos de dados que não são disponibilizados foram contabilizados na categoria *Não Publicado* nesta tabela. Como alguns trabalhos utilizaram uma combinação entre dois ou mais corpus, cada um deles foi contabilizado também, por isso os somatório da quantidade de trabalhos na Tabela 3.4 não é igual ao número de trabalhos analisados.

Tabela 3.4 – Corpus utilizados nos trabalhos avaliados. Os corpus não publicados tiveram seus idiomas classificados como Não se Aplica (NA), mas a Tabela 3.3 indica especificamente cada idioma de cada trabalho.

Corpus	**Quantidade**	**Idioma**
NYT	27	Inglês
Não Publicado	24	NA
SemEval	17	Inglês
Wikipedia	5	Inglês
ACE	3	Inglês
MADE Challenge	2	Inglês
TAC DDI	1	Inglês

Conjuntos de domínio Geral como NYT e Wikipedia são mais utilizados pelos autores, pois são conjuntos de dados semi-estruturados que contêm grande número de registros, ou seja, os pesquisadores não precisam construir ou anotar nenhum registro. Quando o domínio dos dados é mais específico, os autores precisam criar o seu conjunto de dados, o que demanda esforço em reunir e anotar os dados, além de dificultar a comparação de resultados entre esses trabalhos. Também é possível afirmar de acordo com a Tabela 3.3 que há uma concentração de métodos especializados em textos escritos em inglês em face do idioma português.

A Tabela 3.1.7 apresenta os melhores resultados alcançados abordando a tarefa de ER, com o conjunto de dados do NYT, utilizando algoritmos de aprendizado profundo como *Transformers* e redes neurais profundas.

Citação	Ano	Técnica	Precisão
[YCC+20]	2020	RNC + BERT	0,870%
[WXD20]	2020	RNC + RNR - BiLSTM	0,843%
[YDZ+19]	2019	RNR - Bi-GRU	0,826%
[HZS19]	2019	RNC	0,809%
[WLR+19]	2019	Supervisionada à distancia - RNC	0,806%
[ZLWX19]	2019	RNC + RNR - Bi-GRU	0,794%

Tabela 3.5 – Lista dos trabalhos selecionados.

Sobre a comparação de resultados, foi constatado um padrão para as métricas de avaliação, sendo as métricas de *F-Measure*, Precisão e *Recall* as mais utilizadas. De todos os trabalhos revisados, apenas um não utilizou ao menos uma dessas três métricas, resultando em 98,7% de utilização de ao menos uma dessas métricas.

3.2 Trabalhos específicos para a língua portuguesa

Visando complementar a revisão sistemática para encontrar trabalhos que tratem a tarefa de ER para o idioma português, foi feita uma busca manual complementar através do Google Scholar[5]. Assim, também seria possível analisar os resultados de alguns trabalhos que abordem o mesmo idioma que o estudo proposto.

Gamallo e Garcia [GG11] apresentam uma abordagem baseada em regras e em simplificação das estruturas linguísticas para a tarefa de ER. Utilizando um pacote de código aberto de análise sintática multilíngue, *DepPattern*, removeram alguns elementos da entrada realizando uma espécie de compressão da frase para as regras de extração. Esta simplificação permite aplicar regras genéricas de extração semântica, obtidas com uma estratégia de supervisão à distância que aproveita recursos semiestruturados. As regras são adicionadas a uma gramática de dependência parcial, que é compilada em um analisador capaz de extrair instâncias das relações desejadas. Os autores criaram o seu próprio conjunto de dados extraindo frases da Wikipedia para o português e utilizaram apenas relações de duas categorias, profissão e local de nascimento. Segundo os autores, a abordagem não obteve bons resultados ao utilizar correspondência de padrões para classificar as frases, mas obteve bons resultados ao usar regras de dependência.

Também com o intuito de classificar as relações semânticas Batista et al. [BFS+13], utiliza KNN para fazer a classificação, aproveitando um método eficiente baseado em valores mínimos de funções de dispersão como forma de medir a similaridade entre relações, para diferentes tipos de relações semânticas. O método proposto é avaliado na tarefa específica de extração de relações em textos escritos em português, sendo os exemplos

[5]https://scholar.google.com/

de treino extraídos automaticamente da Wikipedia, com base nas relações entre pares de entidades que se encontram explicitamente codificados na DBPedia [6]. Garcia e Gamallo [GG11] optaram consideraram dez tipos diferentes de relações semânticas, oito delas correspondendo a relações semânticas assimétricas e obtiveram resultados de até 55.6% em termos de *F-Measure*.

Taba e Caseli [TC14] em seu estudo, compararam algumas abordagens baseadas em regras com sistemas baseados em aprendizado de máquina para a tarefa de ER, classificando 7 tipos de relações. Dois classificadores de aprendizado de máquina foram investigados neste trabalho, árvores de decisão e *Support Vector Machines*. Diferentes características dos níveis superficial, morfológico e sintático foram definidas para caracterizar os dados de treinamento. Alguns exemplos de características são a distância entre os termos, número de vírgulas entre os termos, classes morfológicas, entre outros. Dois corpus foram utilizados neste trabalho: o primeiro é o CETENFolha [7] e o segundo corpus utilizado foi composto por 646 artigos da *Pesquisa FAPESP* [8], uma revista de divulgação científica. A fim de aplicar os métodos de aprendizagem supervisionada para extrair relações semânticas de textos, uma amostra de ambos os corpora foi anotada manualmente com os termos de interesse e relações entre eles. Os autores destacaram que SVM se comportou muito melhor que Árvores de Decisão e sistemas baseados em regras, principalmente na classificação de relações do tipo "propriedade de".

Diferente dos autores anteriores, Collovini et al. [CMV16b] abordam a tarefa de ER extraindo trechos da sentença que representam a relação semântica e depois as classifica utilizando CRF. Neste trabalho, os autores usam recursos específicos de relação para o português como: *Part-Of-Speech*; léxico (por exemplo, forma canônica), sintático (por exemplo, marca sintática); padrões (por exemplo, verbo seguido por uma preposição); sequência frasal (por exemplo, tags POS da sequência de palavras entre duas ENs); semântica (por exemplo, categoria NE). Como conjunto de dados usaram subconjuntos da Coleção Dourada do HAREM para Reconhecimento de Entidades Nomeadas com anotação manual de qualquer descritor de relação ocorrendo entre pares de ENs na mesma frase do texto. O trabalho contou com poucas amostras mas conseguiu melhores resultados classificando as relações do que extraindo elas.

Em outro trabalho, Collovini et al. [CMV16a] apresentam a extração e estruturação de relações abertas entre entidades nomeadas a partir de textos no idioma português. Os autores aplicaram o mesmo modelo de *Conditional Random Fields* para a extração de descritores de relação entre entidades nomeadas pertencentes às categorias Pessoa, Local e Organização e posterior estruturação dessas relações abertas entre entidades nomeadas no domínio da Organização para o português. Uma pontuação de 0,64 de *F-Measure* foi alcançada como resultado. Como continuidade de seu trabalho, Collovini e Vieira em [CV17]

[6] https://www.dbpedia.org/resources/
[7] https://www.linguateca.pt/cetenfolha/
[8] https://revistapesquisa.fapesp.br/

lançam o sistema de Extração de Relações Abertas para o Português, denominado RelP. O sistema RelP tem como objetivo extrair qualquer descritor de relação que descreva uma relação explícita entre entidades nomeadas no domínio da organização. Para implementar o trabalho, as autoras usaram o esquema de representação, recurso baseado em trabalhos anteriores e um corpus de referência.

Já em seu trabalho mais recente, Collovini et al. [CGC+20] se aproximam bastante ao objetivo do objetivo proposto neste trabalho. Os autores demonstram como o ER pode fornecer subsídios para IC coletando e organizando informações externas de dados não estruturados coletados de jornais, blogs, revistas e portais informativos. Para isso, criaram um *framework* conjunto para identificar as ENs e em seguida extrair as relações semânticas entre essas ENs reconhecidas em um determinado texto. O pré-processamento é responsável pela segmentação e tokenização da frase. O primeiro módulo é responsável por quebrar a sentença e separar todas as palavras, pontuação e marcações que estão dentro dela.

Como parte desse framework, a tarefa de REN também é feita por um modelo preditivo previamente treinado por Santos et al. [dSN14] que usou algoritmo de Rede Neural BiLSTM-CRF. O módulo de recursos é responsável por gerar vetores de recursos para cada par de entidades nomeadas e também para as palavras intermediárias. Por fim, o módulo do modelo CRF é responsável por aplicar o modelo gerado a partir desses vetores de recursos. Neste *framework*, foi utilizado o sistema RelP [CV17], construído para extração de relações abertas para textos em português. O RelP extrai todos os trechos descritores de relação semântica que expressam uma relação explícita que ocorre entre pares de ENs (Organização, Pessoa ou Local) no domínio da organização. O sistema RelP foi aplicado ao corpus composto por frases sobre o mercado financeiro anotado manualmente, totalizando menos de 500 amostras.

Glauber et al. [GCdO19] descrevem a participação dois sistemas com foco na tarefa de ER para textos em português, no Fórum Ibérico de Avaliação de Línguas 2019. Assim, usaram DependentIE e DptOIE, para tratar a tarefa de ER, mais especificamente OpenIE utilizando conjunto de dados concedido pelo Fórum. O primeiro sistema, o DependentIE, é um sistema que usa um Analisador de Dependência para identificar partes úteis de uma frase como sujeito, objetos diretos e indiretos, verbo, advérbio e etc. O segundo sistema, DptOIE é uma evolução do sistema DependentIE e usa *Dependency Parser* de Stanford, regras específicas para extrair fatos em sentenças em português, adapta a pesquisa em profundidade para explorar a árvore de dependências e lida com casos particulares em sentenças com conjunções coordenadas, orações subordinadas e apostos. Os valores para todas as medidas de performance são pouco expressivos. Ainda assim, o DptOIE tem uma pequena vantagem na comparação de ambos os sistemas.

Cabral, Souza e D. Claro apresentam o TabOIEC [CSC20], um classificador multilíngue baseado em características morfossintáticas genéricas. Esse classificador possui um método de caixa de vidro que pode fornecer interpretação sobre algumas das suas de-

cisões. Usaram também um modelo *UDify* para capturar recursos morfo-sintáticos que servem como entrada para modelos classificadores que os autores fizeram testes. Os recursos utilizados foram: distância relativa entre as peças, recursos UPOS, recursos UFeat, Árvore de dependência, Tags e localização do cabeçalho. Durante a etapa de pré-processamento, o objetivo dos autores era converter a saída textual dos extratores OpenIE em um formato estruturado para ser processado nas etapas posteriores. Para comparabilidade, os autores empregaram os mesmos dados usados por Cabral et al. [CGSC20] para seu classificador multilíngue e atingiram até 65% em *F-Measure*.

A combinação de inferência, contexto e intenção permite a extração de fatos implícitos de textos atingindo um primeiro nível pragmático. Assim, Sena e Claro [SC20] apresentam o PragmaticOIE, método de ER que aumenta o número de fatos, extraindo relações de uma frase analisando inferência, contexto e intenção baseada em regras. O módulo inferencial foi retirado de trabalhos anteriores e garante uma interpretação semântica da sentença. Esta abordagem foi composta por regras de transitividade e simetria, considerando regras gerais para inferir novos fatos implícitos. Esse método incorpora aspectos da linguagem e sua estrutura escrita. O PragmaticOIE começa processando as sentenças por meio de um POS *tagger* e um analisador de *chunker* NP. A frase pré-processada passa por uma etapa de extração. Nesta etapa, as relações semânticas são primeiramente identificados, por meio de restrições sintáticas. Em seguida, os fatos são aplicados à etapa de inferência e submetidos a regras transitivas e simétricas. Se algum padrão for encontrado, novos fatos (por meio de transitividade ou simetria) são inferidos. Por fim, o estudo foi avaliado por meio de cinco conjuntos de dados. Duas fontes foram usadas para criar esses conjuntos de dados, Wikipedia e CETENFolha [9].

Diferente da Revisão da Literatura feita na seção 3.1, a pesquisa complementar mostrou que as estratégias para abordar a tarefa de ER para o idioma português ainda dependem de muitos recursos extras como análises de dependência, taggeamentos e POS, como nos trabalhos [GCdO19, CMV16b, GG11]. Existem também trabalhos como [CSC20, SC20] que se baseiam em regras, o que acaba dependendo de um esforço humano muito grande para que se faça um rastreamento de padrões de regras de escrita. Também pode-se observar que existem poucos conjuntos de dados para permitir comparações mais justas entre esses trabalhos, visto que nenhum deles utilizou o mesmo conjunto de dados. Em relação ao trabalho proposto, diferente dos trabalhos revisados, usamos o BERT para obter a semântica das frases sem o uso de seleção aprimorada de recursos ou outros recursos externos. A eficiência da abordagem foi verificada utilizando um conjunto de dados anotado manualmente que também foi disponibilizado para a comunidade. Sobre as métricas de avaliação, foram usadas as métricas padrão dos trabalhos revisados, como *F-Measure*, *Recall* e Precisão, porém com uma adaptação da métrica de Jaccard para assim, ter uma avaliação mais justa da abordagem conjunta construída.

[9] http://www.linguateca.pt/cetenfolha/—versão 2008.

4. ARQUITETURA PROPOSTA

Neste estudo trabalhamos com um escopo no qual uma sentença pode possuir um ou mais pares de ENs e que esses pares de ENs podem conter uma relação semântica entre elas ou não. Acredita-se que esse escopo reflete bem a realidade dos textos não estruturados disponíveis no cotidiano de um setor de IC. Dessa maneira, para extrair as relações semânticas entre esses pares de entidades, propomos uma abordagem *pipeline* que possui 2 modelos de predição: (1) Modelo Classificador apresentado na Seção 4.1; (2) Modelo Extrator de Relação apresentado na Seção 4.2;

A arquitetura proposta pode ser vista resumidamente na Figura 4.1. O primeiro modelo indicará se o par de ENs testado contém ou não uma relação semântica, constituindo um problema de classificação binária. Caso o primeiro modelo indique que o par de ENs testado possui uma relação, o registro é enviado ao segundo modelo para que este extraia os *tokens* que representem essa relação semântica. Essa abordagem permite que os modelos sejam treinados de forma independente e de acordo com o seu propósito, garantindo que possamos efetuar alterações para melhorar o seu desempenho de uma maneira mais facilitada.

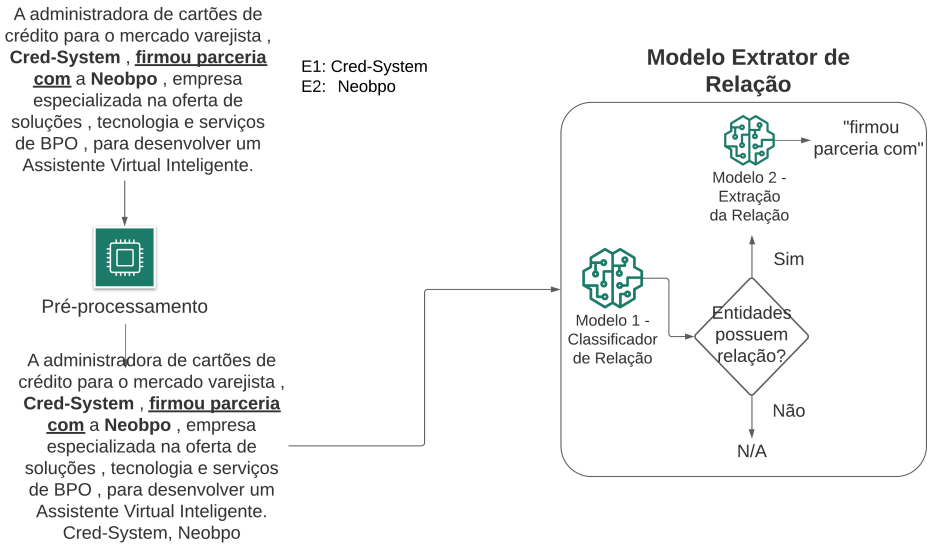

Figura 4.1 – Arquitetura de modelo de extração de relações entre entidades nomeadas.

4.1 Modelo Classificador

Nesta seção, detalhamos nosso modelo classificador baseado em BERT e publicado no *Hackashop on News Media Content Analysisand Automated Report Generation* da *European Chapterof the Association for Computational Linguistics* (EACL) [DLRBVM21].

Conforme mostra a Figura 4.2, ele contém três partes: (1) Camada de entrada; (2) Camada de BERT; e (3) Camada de saída, que é composta por uma função de ativação Sigmóide e dois neurônios que representam as classes a serem previstas.

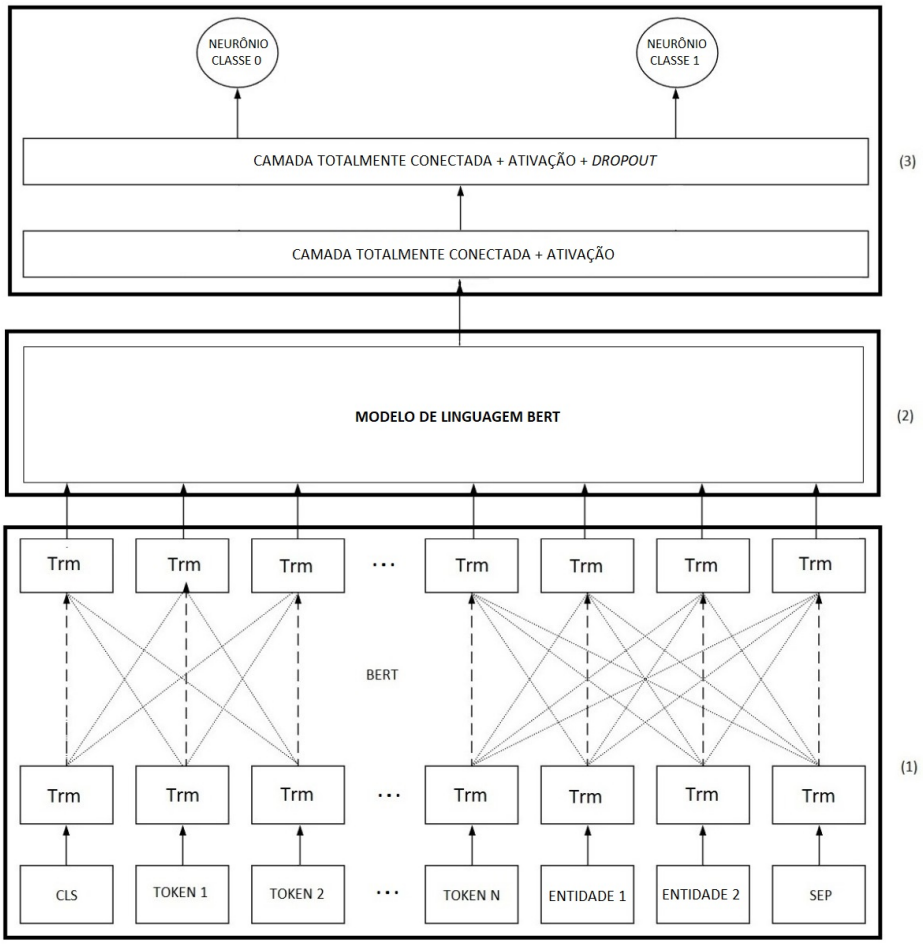

Figura 4.2 – Arquitetura de modelo classificador completa com suas 3 camadas: (1) Camada de entrada; (2) Camada de BERT; (3) Camada de saída.

A camada de entrada consiste em um codificador BERT usado para tokenização da sentença de entrada que produz uma tupla de matrizes (*token*, máscara, ids de sequência), que são usadas como entrada para a segunda camada que é o modelo da linguagem BERT. Utilizamos o modelo pré-treinado em português baseado no corpus brWaC fornecido por Souza et al. [SNL20]. As implementações desta versão de transformadores são fornecidas por Huggingface [1], e o modelo foi construído com as bibliotecas do *Simple Transformers* [2]. A Figura 4.3 ilustra a entrada do modelo proposto que consiste na frase original com as entidades mencionadas e nas entidades a serem verificadas concatenadas. Um *token* especial [CLS] e um *token* [SEP] são adicionados no início e no final da entrada, respectivamente, conforme mencionado na implementação original do BERT [DCLT19]. O

[1] https://github.com/huggingface/transformers
[2] Disponível em https://simpletransformers.ai/

token especial [CLS] é usado para marcar o início do nosso texto. O *token* especial [SEP] é usado para marcar o final de uma frase ou a separação entre duas frases.

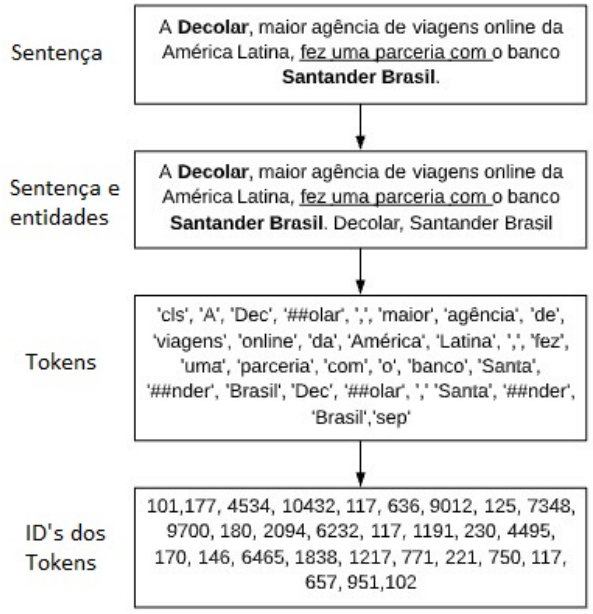

Figura 4.3 – Exemplos de transformações de dados na camada de entrada do modelo. As entidades a serem avaliadas aparecem em negrito e o texto que representa a relação semântica entre elas está sublinhado.

A terceira camada da arquitetura do modelo classificador é identificada como a camada de saída. Esta camada está totalmente conectada com uma função de ativação tangente. A saída desta camada é propagada para uma nova camada totalmente conectada, com função de ativação Sigmóide, cuja característica é o mapeamento dos valores de entrada para 0 ou 1. Neste modelo, esses valores representam não-relação e relação, respectivamente. Conforme mostrado na Figura 4.2, esta camada ainda possui dois neurônios de saída, que indicam as respectivas classes a serem previstas pelo modelo. No final, adicionamos uma camada de *dropout* com uma taxa de 0,1 para evitar *overfitting* do modelo, que acontece quando o modelo memoriza os dados de treinamento e, portanto, perde o poder de generalização.

4.2 Modelo Extrator de Relações

Sentenças da área financeira contêm relações compostas por vários *tokens*, além do fato das frases poderem ser longas. Existem muitas maneiras de expressar relações e a mesma semântica pode ser expressa de diferentes formas. Assim, o problema que abordamos pode se beneficiar do uso de um mecanismo de atenção.

Nesta seção, apresentamos nosso modelo para extração da relação, que também é baseado em *transformers* [VSP+17] BERT e foi aceito para publicação em *Brazilian Con-*

ference on Intelligent Systems 2021 (BRACIS) [GTM+21]. Utilizamos a mesma versão de *transformers* que no modelo de classificação, fornecida por Huggingface, e adicionamos e treinamos as últimas camadas do modelo para extrair uma subsequência que representa a relação entre as entidades nomeadas do texto de entrada. Para a implementação da rede neural, a biblioteca TensorFlow Keras [3] foi usada para adaptar a última camada de BERT.

Como o modelo de extração é baseado em *transformers*, ele é composto por dois componentes principais, *encoder* e *decoder*, como ilustrado pela Figura 4.4. Essa abordagem fornece uma representação ligeiramente diferente para as palavras à medida que são usadas nas frases e sua relação com outras palavras. Isso permite que o modelo compreenda a própria palavra e o contexto no qual está inserida.

Em termos gerais, as sentenças entram através do *encoder*, que possui duas camadas: um mecanismo de atenção que então se alimenta em uma *Rede Neural Feed Forward*; a saída do *encoder* é, então, enviada para o *decoder*. O *decoder* também tem essas duas camadas, mas entre elas há adicionalmente um mecanismo de atenção que ajuda a focar nas partes relevantes da frase de entrada. A seguir, descrevemos com mais detalhes alguns desses conceitos.

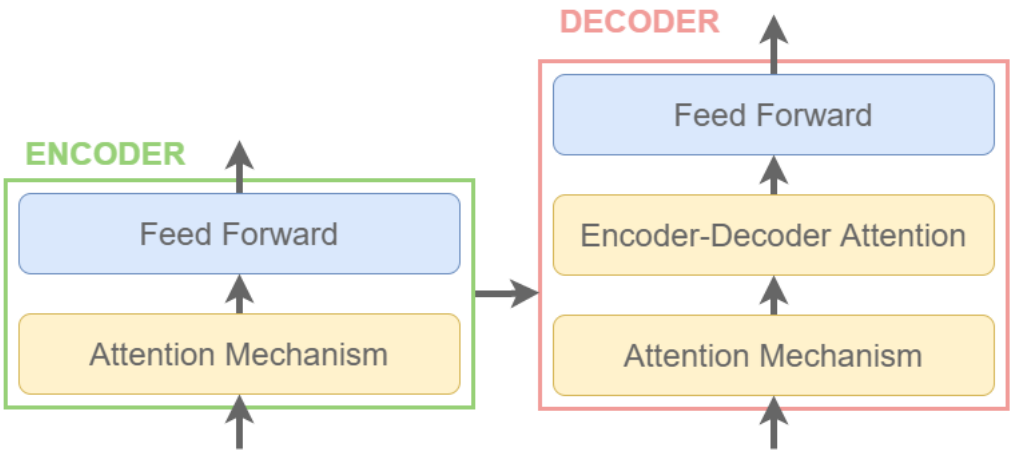

Figura 4.4 – Arquitetura do modelo extrator de relações.

***Attention Mechanism*:** Um mecanismo de atenção pode ser descrito como o mapeamento de uma consulta e um conjunto de pares de valores-chave para uma saída, sendo que a consulta (Q), as chaves (K), os valores (V) e a saída são todos vetores. De acordo com a Equação 4.1 de Vaswani et al. [VSP+17], a saída é calculada como uma soma ponderada de valores, onde o peso (W) atribuído a cada valor é calculado por uma função de correspondência de consulta com a chave correspondente. No modelo pré-treinado de BERT é usado *Multi-Head Attention*, o que permite que o modelo utilize em conjunto, informações de diferentes representações em diferentes posições.

[3] https://www.tensorflow.org/

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O, \tag{4.1}$$

onde

$$head_i = Attention(QW_1^Q, KW_i^K, VW_i^V). \tag{4.2}$$

Podemos obter representações de palavras mais ricas usando mecanismo de atenção. Essas representações de palavras são consideradas como o contexto, com base na correlação entre as palavras em uma frase [LSC19]. Um mecanismo de atenção funciona comparando cada palavra na frase com todas as demais palavras na frase, incluindo a si mesma, e pesando novamente as representações de cada palavra para incluir relevância contextual. Em outras palavras, o mecanismo de atenção permite que as entradas interajam umas com as outras e descubram a quem devem prestar mais atenção. Por exemplo, na frase, *"Linx foi adquirida pelo Banco Stone"*, *Linx* seria comparada com *Linx*, *foi*, *adquirida*, *pelo*, *Banco* e *Stone*, e como *"Linx"* é comparada com essas outras palavras, sua representação de palavras seria recalculada para incluir a relevância das palavras para seu próprio significado na frase em conformidade.

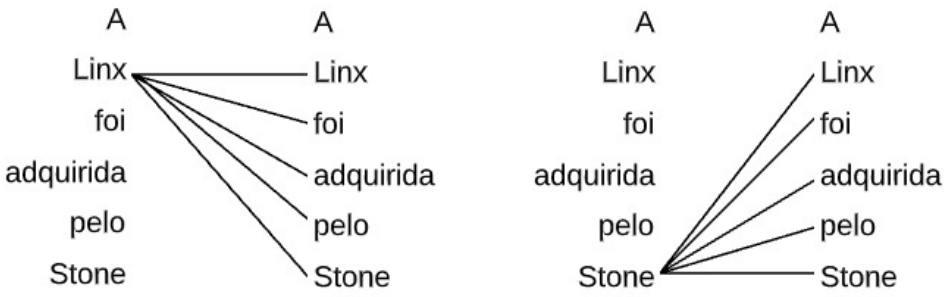

Figura 4.5 – Visualização do mecanismo de atenção.

A Figura 4.5 ilustra os resultados do mecanismo de atenção na frase, *"A Linx foi adquirida pelo Stone"*, que possui uma relação semântica de aquisição entre *Linx (e1)* e *Stone (e2)*. Na Figura 4.5, o lado esquerdo representa as palavras em que *Linx*, a primeira entidade, se concentra e o direito representa as palavras em que *Stone*, a segunda entidade, se concentra. Podemos reconhecer que o par de entidades está comumente concentrado em *foi*, *adquirida*, *pelo* e uma na outra, destacando assim, a relação semântica entre o par de ENs.

Camada *Feed-Forward*: Além do mecanismo de atenção, cada uma das camadas em nossos *encoders* e *decoders* contém uma camada *feed-forward* totalmente conectada, que é aplicada a cada posição separadamente e identicamente. Isso consiste em duas transformações lineares que são calculadas de acordo com a Equação 4.3, com uma função de ativação ReLU entre elas [VSP+17].

$$FFN(x) = max.(0, xW_1 + b_1)W_2 + b_2. \tag{4.3}$$

5. EXPERIMENTOS

Neste capítulo descrevemos os experimentos realizados no corpus de informações do domínio financeiro para demonstrar a eficácia da abordagem proposta para extração das relações entre entidades nomeadas e analisar seu desempenho. O estudo proposto é detalhado seguindo a metodologia clássica de *Knowledge Discovery in Databases* - KDD, proposta por Fayyad em [FPSS96]. Conforme ilustra a Figura 5.1, este processo contém cinco etapas que vão desde a coleta e criação do corpus, até a avaliação dos resultados. As próximas seções visam explicar como cada uma dessas etapas foi empregada neste trabalho.

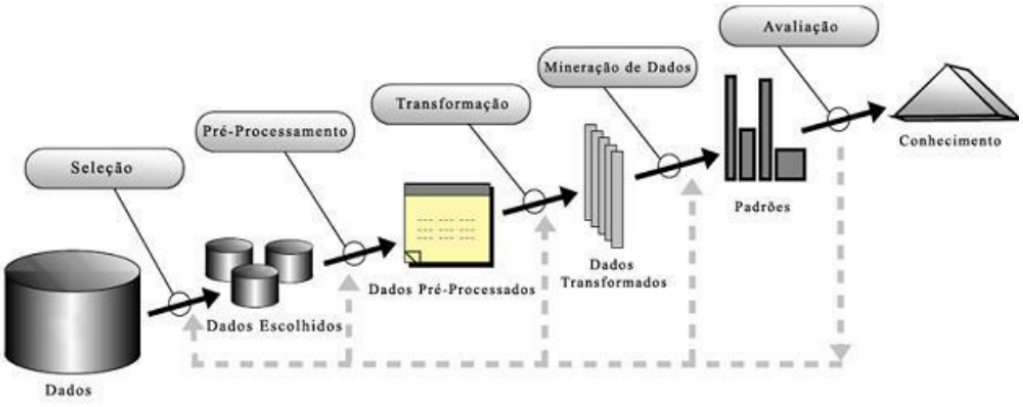

Figura 5.1 – Metodologia de Descoberta de Conhecimento em Bases de Dados proposta por Fayyad et al. [FPSS96].

5.1 Seleção

Como na etapa de seleção é necessário indicar quais dados serão utilizados durante os experimentos para a tarefa ER [FPSS96], iniciamos pela busca por corpora. No entanto, não houve evidências de conjuntos de dados abertos no contexto de extração de relações na área financeira em língua portuguesa. Portanto, para este trabalho, foi criado um corpus com 9.114 tuplas *(e1, e2, relação)* anotadas manualmente.

Para compor este corpus, inicialmente foram coletadas mais de 1.500 notícias do mercado financeiro a partir de 2018, que foram fornecidas por uma empresa parceira. Essas notícias foram coletadas em vários meios de comunicação, como sites do mercado financeiro, jornais e balanços corporativos.

Outros 7.097 *tweets* com datas a partir de janeiro de 2021 também foram coletados. Esses *tweets* foram selecionados de 10 usuários de meios de comunicação com foco no mercado financeiro e na economia em geral. Os usuários selecionados foram: *@info-*

money, *@EstadaoEconomia*, *@UOLEconomia*, *@g1economia*, *@OGlobo_Economia*, *@folha_mercado*, *@InvestingBrasil*, *@leiamoneytimes*, *@valoreconomico*, *@br_economico*.

Cada uma dessas notícias e *tweets* foram separadas de acordo com as suas sentenças. Dessa maneira, esse material deu origem a mais de 10.000 frases que foram analisadas manualmente. As frases que continham co-referências foram removidas porque lidar com elas exigiria processamento adicional enquanto que, as frases restantes passaram para a próxima fase da metodologia KDD. A Tabela 5.1 ilustra alguns exemplos de frases e suas anotações.

Sentença	**Entidade_1**	**Relação**	**Entidade_2**
Conselheiros do Banco do Brasil pedem manutenção de Brandão no cargo de CEO.	Conselheiros do Banco do Brasil	pedem manutenção de	Brandão
CrediPronto é a especialista em imobiliário do Itaú e nova parceira da Franq.	CrediPronto	nova parceira da	Franq
PicPay, da família Batista, se prepara para lançar ações nos Estados Unidos.	PicPay	se prepara para lançar ações nos	Estados Unidos
A Climatempo, maior empresa privada de meteorologia do país , acaba de anunciar sua parceria com o Banco do Brasil.	Climatempo	acaba de anunciar sua parceria com o	Banco do Brasil
Veja as análises para Gol, Azul e Latam	Gol		Azul
Após vender pedaço do Shopping Leblon, Brookfield procura comprador para o Rio Sul.	Shopping Leblon		Rio Sul

Tabela 5.1 – Exemplos de Tuplas selecionadas para o corpus.

5.2 Pre-processamento

A etapa de pré-processamento destina-se a aplicar procedimentos de limpeza, correção ou remoção de dados inconsistentes ou ausentes obtidos ao final da etapa de Seleção 5.1. Dessa forma, executamos um processo de correção ortográfica, padronização de acrônimos e extensão de siglas, limpeza de caracteres especiais, identificação de entidades nomeadas e remoção de sentenças sem entidade nomeada ou com apenas uma entidade nomeada.

No contexto deste trabalho, realizou-se um processo de verificação ortográfica manual para cada frase, utilizando o corretor ortográfico do Excel. Acrônimos também foram expandidos, como *BC* sendo substituído por *Banco Central*. Este tipo de padronização foi feita manualmente para o corpus deste trabalho, percorrendo cada linha do conjunto de da-

dos via Excel, mas entendemos que em um cenário de trabalho real, essa tarefa se torna enorme e pode ser automatizada criando uma base de entidades nomeadas e seus acrônimos. Assim, é possível desenhar um processo que valide as siglas da frase e as substitua por extensões ou mesmo uma abordagem que enfoque apenas entidades específicas informadas pelo analista de IC. O processo de limpeza dos dados foi feito automaticamente por meio de um *script* disponível no Github [1] que remove caracteres especiais e siglas que seguem a própria descrição.

As entidades nomeadas foram identificadas por meio de uma única ferramenta para Reconhecimento de Entidades Nomeadas (REN), denominada SpaCy [2], garantindo que os mesmos critérios sejam utilizados para todas as sentenças. As frases que têm apenas uma entidade nomeada reconhecida ou nenhuma são removidas do conjunto de dados. Ao final desta etapa de limpeza, pouco mais de 5.000 frases contendo duas ou mais ENs foram selecionadas para a próxima etapa. Essas ENs em questão são aquelas relacionadas às categorias pessoa, lugar e organização, pois foram consideradas as mais relevantes para o domínio das organizações [CGS+11]. O ponto focal são informações sobre organizações, bem como seus relações com outras organizações, pessoas e lugares.

5.3 Transformação

A etapa de transformação ou formatação dos dados analisa os dados obtidos na etapa anterior e os reorganiza de forma específica para que possam ser analisados e interpretados pelo algoritmo na etapa seguinte. No caso deste estudo, após a identificação das Entidades Nomeadas na fase anterior, o primeiro passo da etapa de transformação foi combinar todas as entidades presentes na frase para criar uma tupla (frase, entidade, entidade) para cada combinação. É importante ressaltar que sentenças com mais de duas entidades geraram mais de uma tupla para a mesma frase, pois foi necessário criar uma tupla para cada combinação de entidades.

Em seguida, as relações semânticas entre as entidades nomeadas destacadas foram anotadas manualmente. Uma tupla foi considerada positiva quando existe qualquer relação semântica entre duas entidades nomeadas das categorias definidas na Seção 5.2. Quando não há esta relação, a tupla foi considerada negativa.

Após o término da anotação manual das relações entre as entidades, o conjunto contendo pouco mais de 5.000 sentenças obtido na etapa anterior de Pré-processamento, deu origem ao corpus final, que é composto por 9.114 registros. Desse total, 4.641 (50.9 %) são tuplas positivas, ou seja, contém uma relação entre as entidades destacadas, e 4.473 (49.1 %) são tuplas negativas, onde não há relação entre as entidades. Finalmente,

[1] https://github.com/DanielReeyes/financial-market-corpus
[2] https://spacy.io/

as duas entidades nomeadas são concatenadas no final da frase. A Tabela 5.2 exemplifica alguns registros que possuem combinações de entidades que podem gerar mais de uma tupla por frase e também exemplos de tuplas com anotações positivas que contêm relações semânticas entre entidades nomeadas do tipo de organização.

Instancia	**Relação Semântica**
O estudo na britânica Nature é de autoria de **Neil M. Ferguson**, do **Imperial College**, de Londres, e mais sete colaboradores.	do
O estudo na britânica Nature é de autoria de Neil M. Ferguson, do **Imperial College**, de **Londres**, e mais sete colaboradores.	de
Rappi faz parceria com a **Visa** e anuncia cartão pré-pago no Brasil.	faz parceria com a
Rappi faz parceria com a Visa e anuncia cartão pré pago **Brasil**.	anuncia cartão pré-pago no

Tabela 5.2 – Exemplos de tuplas positivas com anotações mostrando as relações entre entidades nomeadas. As entidades a serem avaliadas aparecem em negrito e o texto que representa a relação semântica entre elas está sublinhado.

As sentenças são naturalmente compostas por palavras e caracteres, a etapa de transformação da metodologia também incluiu a transformação de *tokens* em representações numéricas pelo codificador BERT. Conforme descrito na Seção 4.1, o BERT adicionou os *tokens* especiais [CLS] e [SEP] devidamente codificados em cada frase, finalizando a transformação da frase em linguagem natural na entrada para o extrator de modelo.

5.4 Mineração

A etapa de mineração contempla a tarefa de previsão, na qual um padrão comportamental dos dados é pesquisado a fim de prever o comportamento de uma entidade futura [FPSS96]. O corpus foi dividido aleatoriamente em duas partes: um conjunto com 90% das tuplas que foram utilizadas para treinar o modelo; e outro conjunto com 10% das tuplas para teste; Optou-se por essa taxa de proporção para que fosse possível disponibilizar mais amostras para treinamento aos dois modelos utilizados na abordagem.

Assim, após a etapa de treinamento, a partir da qual o modelo é capaz de reconhecer esse padrão, foi possível aplicá-lo aos dados do conjunto de testes. De acordo com a Tabela 5.3, cada conjunto manteve o balanceamento original do conjunto de dados, neste caso, tuplas com relação semântica (positiva) e sem relação semântica (negativa).

O ajuste dos hiper-parâmetros de BERT usados em cada um dos modelos descritos no Capítulo 4 foi feito usando a combinação de todos os valores indicados por Jacob

Conjunto	Amostras	Distribuição Classe Positiva (%)	Amostras Positivas
Original	9114	50.9	4641
Treinamento	8203	50.9	4177
Teste	911	50.9	464

Tabela 5.3 – Composição de cada conjunto de dados utilizado nos experimentos.

Devlin quando ele propôs o novo modelo da linguagem BERT em [DCLT19]. Nesse trabalho, ele usou a maioria dos hiper-parâmetros com valores padrão, exceto para tamanho do lote, taxa de aprendizado e um número de épocas de treinamento. A taxa de *Dropout* sempre foi mantida em 0,1. Assim, os valores analisados para a tarefa deste trabalho foram:

- **Batch:** 16, 32;
- **Taxa de Aprendizado:** 2e-5, 3e-5, 5e-5;
- **Períodos:** 2, 3, 4, 5;
- **Otimizador:** AdamW;

No final, para cada um dos modelos que fazem parte da abordagem proposta, executamos um total de 24 experimentos com todas as combinações possíveis dos parâmetros descritos acima. Após a análise dos resultados, os modelos foram definidos considerando os parâmetros que apresentaram os melhores resultados e estão descritos na Tabela 5.4.

	Modelo Classificador	Modelo Extrator
Hiper-parâmetro	**Valor**	**Valor**
Batch	32	32
Taxa de Aprendizado	5e-5	2e-5
Períodos	4	5
Otimizador	AdamW	AdamW

Tabela 5.4 – Combinação de hiper-parâmetros que apresentou os melhores resultados.

5.5 Avaliação

A última etapa da metodologia KDD apresentada neste capítulo visa avaliar o desempenho do modelo que construímos. A avaliação experimental foi realizada aplicando, nos dados de teste, os modelos construídos na fase de aprendizagem com base nos parâmetros definidos após o ajuste dos hiper-parâmetros conforme descrito na Seção 5.4.

Conforme a revisão da literatura realizada e apresentada no Capítulo 3 métricas como Acurácia, *Recall*, Precisão, *F-Measure* são comumente utilizadas para avaliar sistemas de ER, e por tal motivo foram empregadas para avaliar o modelo de classificação. A métrica de Coeficiente de Similaridade de Jaccard 2.6 foi aplicada para avaliar o modelo de extração de relação. Essa métrica foi escolhida pelo fato de que pode apresentar uma visão mais realista da eficiência do modelo ao reconhecer os *tokens* presentes na relação. Se considerássemos apenas extrações completamente corretas, o modelo seria penalizado demasiadamente por não extrair apenas um *token*, por exemplo.

Durante a revisão não encontramos trabalhos que utilizassem tuplas positivas e negativas no mesmo conjunto de dados e que o propósito fosse extrair as relações quando essas tuplas fossem positivas. Então, além de avaliar cada um dos modelos separadamente, propomos também uma forma de avaliar a eficiência de toda a abordagem, ou seja, avaliar a extração final da abordagem composta pelos dois modelos. A avaliação consiste em utilizar o Coeficiente de Jaccard em conjunto com as métricas de *Recall*, Precisão e *F-Measure* que são extraídas a partir da contagem dos VP, VN, FP e FN conforme explicado em 2.6.

Ao indicar que há uma relação em uma tupla que realmente contém, ou seja, um Verdadeiro Positivo (VP), essa mesma tupla é enviada para o modelo de extração que tem sua extração avaliada pela métrica de Jaccard. Assim, o valor referente a este VP será o valor correspondente à métrica de Jaccard do modelo extrator, podendo variar de 0 a 1. A partir dessa alteração de validar o VP, podemos calcular as métricas de de *Recall*, Precisão e *F-Measure* de acordo com as suas equações. A Tabela 5.5 ilustra a utilização dessa forma de avaliar a abordagem completa.

Na Tabela 5.5, a primeira instância foi corretamente classificada e extraída em sua totalidade, e o seu valor de VP é 1. Os segundo e terceiro registros foram classificados corretamente mas o modelo extrator identificou apenas uma parte das relações e foi penalizado conforme as pontuações Jaccard de cada um dos registros. Dessa maneira o valor de VP de cada um dos registros é 0,833 e 0,666 respectivamente. O quarto registro não possui relação entre as duas entidades nomeadas testadas, porém a abordagem classificou a tupla como positiva, contabilizando um Falso-Positivo (FP). Por fim, o quinto registro foi classificado como negativo, porém há uma relação semântica entre as entidades nomeadas, caracterizando um erro do modelo e contabilizando um Falso-Negativo (FN). Portanto, ao calcular as métricas da abordagem conjunta para os exemplos listados na Tabela 5.5, chegaríamos ao seguinte resultado utilizando as equações de *Recall*, Precisão e *F-Measure*:

$$Recall = \frac{VP}{VP + FN} = \frac{2,499}{2,499 + 1} = 0,714, \tag{5.1}$$

Tupla	Classificação	Extração	Pontuação Jaccard
A equipe do Jornal do Nikkey **esteve pesquisando um lado gostoso da** Liberdade em São Paulo. Liberdade , Jornal do Nikkey	Possui	esteve pesquisando um lado gostoso da	1
A Credicard **empresa de soluções de pagamento do** banco Itaú Unibanco acaba de anunciar sua parceria com a Chubb Seguros. Credicard , Itaú Unibanco	Possui	empresa de soluções de pagamento do banco	0,833
FMI **destaca recuperação de** EUA e China como motores da economia mundial. FMI , EUA	Possui	destaca recuperação	0,666
Em especial, no caso do BV, a situação do BB, que é um dos sócios e teve recentemente a troca de presidência um fator que gerou tensão. BB , BV	Possui	NA	NA
Folha e IOB **iniciam tira-dúvidas de leitores sobre o** Imposto de Renda. IOB , Imposto de Renda	Não possui	NA	NA

Tabela 5.5 – Exemplo de tuplas e suas classificações e extrações para avaliação da abordagem completa. Em negrito encontram-se destacadas as relações semânticas entre as entidades nomeadas, quando há.

$$Precisao = \frac{VP}{VP + FP} = \frac{2,499}{2,499 + 1} = 0,714, \tag{5.2}$$

$$F\text{-}Measure = \frac{2 * Precisao * Recall}{Precisao + Recall} = \frac{2 * 0,714 * 0,714}{0,714 + 0,714} = 0,714. \tag{5.3}$$

6. RESULTADOS

Neste capítulo serão apresentados os resultados obtidos durante os experimentos de acordo com a metodologia detalhada no Capítulo 5. Durante a Seção 6.1 são apresentados os resultados separados do modelo de classificação binária; Na Seção 6.2 indicamos a pontuação de Coeficiente Jaccard obtida pelo Modelo de Extração; Seção 6.3 contém a avaliação da abordagem *pipeline* completa, contendo os dois modelos. Por fim, a Seção 6.4 há uma discussão acerca dos resultados obtidos assim como uma análise dos erros identificados.

6.1 Modelo de Classificação

Após a etapa de treinamento do modelo, o modelo de classificação foi aplicado ao conjunto de dados de teste. O modelo obteve bons resultados nesta etapa de avaliação, alcançando uma precisão geral e *F-Measure* de 88%. Uma observação importante a se fazer é que os resultados também são bons quando se trata da classe alvo, ou seja, quando o rótulo é positivo, como pode ser visto na Tabela 6.1.

Métrica	**Positiva**	**Negativa**	**Geral**
Precisão	0,88	0,89	0,885
Recall	0,90	0,87	0,884
F-Measure	0,89	0,88	0,884
Acurácia	-	-	0,88

Tabela 6.1 – Precisão, *Recall* e *F-Measure* calculados para cada classe e precisão e *F-Measure* geral do modelo.

Apesar disso, mostra-se que o modelo proposto foi capaz de reconhecer padrões e indicar quando duas entidades estão semanticamente relacionadas em uma mesma frase no domínio financeiro. Em relação ao *Recall*, o experimento indica que, o modelo proposto obteve um desempenho muito bom de aproximadamente 90% quando se trata da classe positiva (tem uma relação). Ou seja, quando realmente pertence à classe positiva, em aproximadamente 90% dos casos, identifica-se corretamente.

6.2 Modelo de Extração

Após a etapa do modelo classificação, 417 amostras que foram corretamente classificadas contendo uma relação semântica entre as ENs testadas, seguiram para a etapa de extração dos *tokens* que expressam essa relação semântica. O Modelo de Extração

também obteve bons resultados, indicando uma métrica Jaccard geral de 76,3% para todos os casos de predição. Em termos de relações extraídas corretamente, foi possível extrair 228 relações completamente corretas de um total de 417 amostras disponíveis para teste; extrações completamente corretas é quando a relação extraída é exatamente a mesma que aquela anotada manualmente. Consideramos uma relação parcialmente correta quando a relação extraída atinge uma pontuação de Jaccard de pelo menos 50%. Quanto às relações parcialmente corretas, nossa abordagem foi capaz de extrair 121 relações. Em apenas 34 ocorrências o modelo não foi capaz de extrair nenhum *token* presente na relação a ser extraída. Estes resultados são apresentados de forma sucinta na Tabela 6.2.

Número de Relações	**Pontuação Jaccard**	**% Totalmente Correta**	**% Parcialmente Correta**	**% Totalmente + Parcialmente Corretas**
417	76,3%	54,7%	29%	83,7%

Tabela 6.2 – Resultados obtidos pelo modelo de Extração.

Avaliação Complementar: Um experimento complementar foi feito para verificar a eficácia do modelo extrator considerando todas as 464 amostras positivas do conjunto de dados de teste. Ao considerar todas as amostras, o modelo manteve os bons resultados, indicando uma métrica Jaccard geral de 76,6% para todos os casos de predição. Em termos de relações extraídas corretamente, foi possível extrair 251 relações completamente corretas de um total de 464 amostras disponíveis para teste; Quanto às relações parcialmente corretas, nossa abordagem foi capaz de extrair 138 relações. Em 75 ocorrências o modelo não foi capaz de extrair nenhum *token* presente na relação a ser extraída. Dessa maneira, é possível inferir que o Modelo de Extração conseguiu manter um padrão em suas extrações e que, mesmo com todas as amostras positivas disponíveis não afetaria em grande escala, o resultado da abordagem conjunta.

6.3 Avaliação Conjunta

Com o objetivo de avaliar o resultado final da abordagem *pipeline*, avaliamos os resultados de acordo com as métricas estabelecidas na Seção 5.5. O modelo de classificação conseguiu classificar corretamente 389 tuplas como negativas de 447 possíveis, enquanto que de 464 tuplas positivas, 417 foram classificadas corretamente como positivas. Assim, a Tabela 6.3 representa os resultados da primeira etapa da abordagem. Portanto, apenas estas 417 amostras foram utilizadas para a segunda etapa, referente à extração dos *tokens* que representam a relação semântica entre as ENs. Essas são as amostras que foram corretamente classificadas como contendo uma relação entre as ENs na primeira etapa.

		Resultado Real	
		1 - Contém relação	0 - Não contém relação
Resultado Predito	1 - Contém relação (+)	417 (VP)	58 (FP)
	0 - Não contém relação (-)	47 (FN)	389 (VN)

Tabela 6.3 – Matriz de Confusão

Conforme dito na Seção 6.2 o modelo de extração alcançou uma pontuação média do Coeficiente de Jaccard de 76,3% entre as 417 amostras disponibilizadas como teste. O somatório obtido de cada pontuação Jaccard para cada tupla de teste foi de 318,05 pontos. Para podermos determinar a eficiência da abordagem conjunta, essa pontuação é indicada como total de VP para assim, podermos calcular as métricas de *Recall*, precisão e *F-Measure*. As equações 6.1, 6.2, 6.3 indicam o resultado geral da abordagem *pipeline* adotada neste estudo.

$$Recall = \frac{318,05}{318,05 + 47} = 0,871. \tag{6.1}$$

$$Precisao = \frac{318,05}{318,05 + 58} = 0,845. \tag{6.2}$$

$$F\text{-}Measure = \frac{2 * 0,871 * 0,845}{0,871 + 0,845} = 0,858. \tag{6.3}$$

6.4 Análise de Resultados

A abordagem proposta obteve bons resultados na etapa de avaliação. A Tabela 6.4 apresenta os resultados parciais para cada etapa da abordagem e a métrica geral dos casos de predição. De acordo com a tabela, a abordagem proposta conseguiu reconhecer em grande parte dos casos, quando uma tupla possui ou não uma relação semântica, fazendo com que uma maior parcela dos dados de teste fosse utilizado para a segunda etapa da ER. O modelo de extração também alcançou boas métricas ao identificar os *tokens* que representam a relação semântica entre as ENs testadas visto que, em aproximadamente 84% do conjunto de testes utilizado na segunda etapa, o modelo conseguiu extrair pelo menos metade da relação semântica.

Um diferencial da arquitetura proposta é trabalhar com sentenças que não possuem relação semântica, visto que grande parte dos trabalhos revisados supõe que sempre

Métrica	Modelo Classificação	Modelo Extração	Abordagem Completa
Recall	88,4%	-	87,1%
Precisão	88,5%	-	84,5%
F-Measure	88,4%	-	85,8%
Jaccard	-	76,3%	76,3%

Tabela 6.4 – Resultados obtidos por cada modelo e resultados gerais.

há uma relação entre as ENs ou consideram que a tupla que não possui relação semântica é uma categoria a ser classificada. Essa proposta foi feita por se assemelhar mais com o cenário atual das notícias de IC. Assim, o modelo de classificação desempenha uma importante função de filtrar essas sentenças e, de acordo com os resultados, também conseguiu reconhecer grande parte desses casos.

Em geral, as relações extraídas expressam informações relevantes para o foco de IC, como relações de formação de parceria entre Organizações; relações de investimento financeiro entre organizações; relações entre Pessoas e Organizações; entrada de organizações no mercado em determinadas localidades. A Tabela 6.5 ilustra instâncias de relações extraídas que estão completamente corretos. A partir dessa abordagem de ER, é possível criar uma base de conhecimento histórica, facilmente acessível e interpretada por analistas de IC. Então, é possível fornecer informações para que todos possam extrair valor de forma rápida e clara.

Sentença	Entidades	Relação
1. A B2W dona da Americanas.com e do Submarino estuda separar o braço da Ame Digital de suas operações.	B2W, Ame Digital	estuda separar o braço da
2. Bolsonaro volta a criticar pesquisa de emprego do IBGE.	Bolsonaro, IBGE	volta a criticar pesquisa de emprego do
3. O Banco BMG e o Clube Atlético Mineiro se uniram para marcar a história mineira mais uma vez com o lançamento do Meu Galo BMG.	Clube Atlético Mineiro, Banco BMG	se uniram
4. O Brasil é um país foco para o Google com alto nível de engajamento diz diretor de dispositivos para a América Latina.	Google , Brasil	é um país foco para o

Tabela 6.5 – Relações completamente corretas extraídas pelo modelo proposto.

A Figura 6.1 ilustra a pontuação média de acordo com o comprimento da relação procurada. De acordo com a Figura, afirmamos que o modelo foi capaz de responder bem à complexidade do tamanho da relação a ser extraída. Também foi capaz de reconhecer

padrões e indicar quais *tokens* pertencem às relações semânticas contidas em uma mesma frase no domínio financeiro.

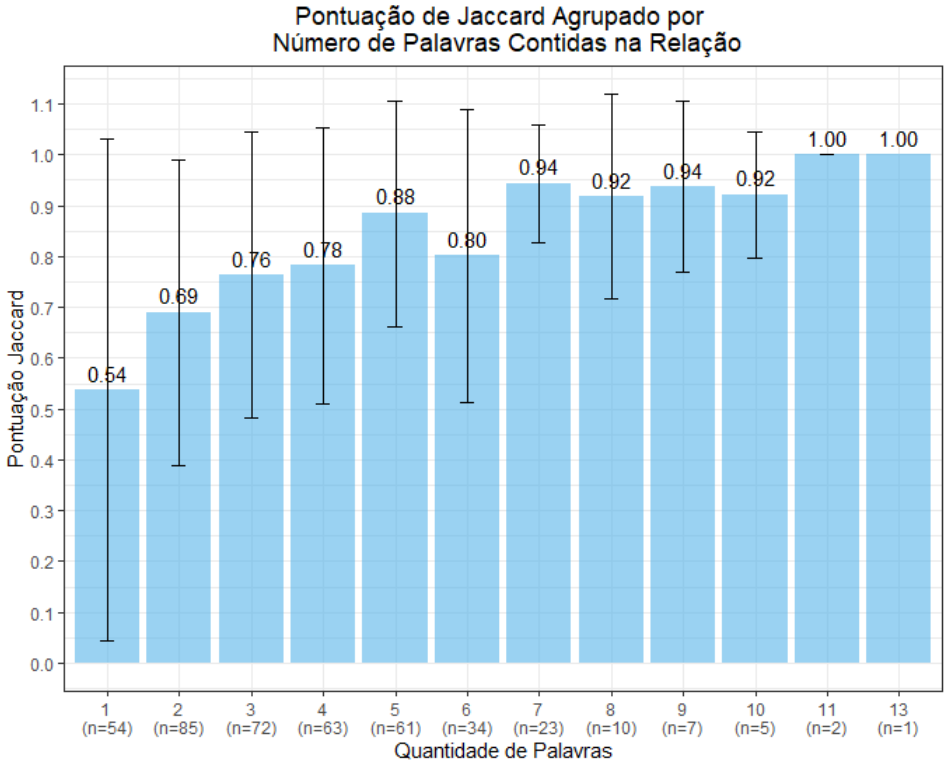

Figura 6.1 – Pontuação de Jaccard média e desvio padrão agrupados pelo número de palavras contidas na relação a ser extraída após correta classificação pelo Modelo Classificador. Neste caso, "n"indica o numero de amostras.

Pode-se inferir também que o modelo teve maior dificuldade ao tentar extrair relações compostas por apenas um *token*, o que acaba penalizando o seu desempenho. Nesses casos, o modelo teve mais dificuldade em inferir relações compostos apenas por uma preposição (por exemplo, da, de, na, no) como mostrado na Tabela 6.6. Este tipo específico de relação semântica possui muitas amostras dentro do conjunto teste e, na maioria dos casos, há apenas duas possibilidades de pontuação, 0 ou 1, quando ele erra toda a extração ou quando ele acerta toda a extração respectivamente. Esse tipo de relação também dificulta fazer uma análise do impacto das *stopwords*, que são palavras que podem ser consideradas irrelevantes para o conjunto de resultados a ser exibido em uma busca realizada. Nesse caso as preposições que compõem as relações semânticas não podem ser removidas por serem exatamente o trecho da sentença a serem extraídas.

Outro ponto identificado é que, conforme mostra a Figura 6.2 o modelo proposto apresentou maior probabilidade de errar no final das relações a serem extraídas. Neste caso, o modelo deixou de extrair os *tokens* finais ou extraiu alguns a mais que não pertenciam à relação semântica. Ainda sobre esse ponto, foi observado que quando o modelo erra no final da relação, em grande parte dos casos é pelo fato de não selecionar uma preposição (por exemplo, da, de, na, no). Em contrapartida, o modelo quando erra ao extrair

Sentença	Entidades	Relação Extraída
1. Nubank chama Daniel Goldberg **da** Farallon para seu conselho de administração.	Farallon , Daniel Goldberg	*None*
2. Doria ataca intervenção de Bolsonaro **na** Petrobras desnecessário e condenável.	Petrobras , Bolsonaro	*None*
3. Com sinal verde para venda de estatais Guedes e Onyx **disputam** programa de privatização do governo.	Onyx , Guedes	disputa
4. O Mercado Pago **fintech do grupo** Mercado Livre e a Hub Fintech startup fornecedora de soluções de negócios em meios de pagamento firmaram parceria para criação de uma oferta compartilhada.	Mercado Livre, Mercado Pago	fintech

Tabela 6.6 – Exemplos de relações não extraídas pelo modelo. Em negrito e sublinhado, a relação semântica a ser extraída.

o início da relação, palavras com maior sentido são suprimidas ou adicionadas. Também há circunstâncias em que o modelo extrator errou tanto no início quanto no final, deixando de selecionar algum *token* ou selecionando *tokens* a mais, este tipo de acontecimento foi o segundo que mais ocorreu e, segundo as análises feitas segue o padrão dos cenários anteriores. Casos contabilizados como *NA* compreendem extrações nas quais toda a relação foi encontrada, ocasiões que nenhuma parte da relação foi extraída.

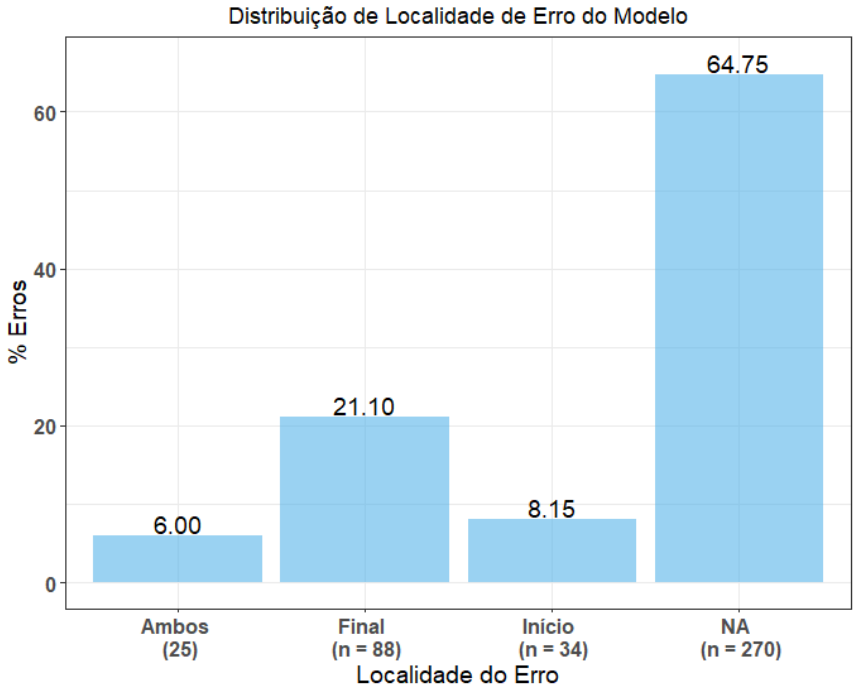

Figura 6.2 – Distribuição de erros de acordo com a localidade da extração no qual ocorreu. Neste caso, "n"indica o numero de amostras e NA indica casos nos quais a análise não se aplica.

7. CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, fornecemos uma abordagem da tarefa de ER e um corpus anotado manualmente a partir de notícias fornecidas por uma empresa de inteligência de mercado e também notícias coletadas no Twitter sobre o tema Inteligência Competitiva. Mais de 1.500 notícias sobre o mercado financeiro e mais de 7.000 *tweets* foram selecionados para construção da base de conhecimento. O modelo demonstrou ser capaz de reconhecer relações semânticas e, portanto, é útil para a descoberta de eventos e fatos relacionados ao mercado financeiro.

A partir da Revisão Sistemática da Literatura apresentada no Capítulo 3, é possível notar que existem poucas pesquisas sobre técnicas de extração de relações entre entidades nomeadas para o domínio financeiro em português. Esse domínio carece de soluções práticas, em parte devido à grande quantidade de informações na área financeira, de forma que a análise manual torna-se difícil para atender às necessidades e fazer pleno uso dessas informações. Devido a essa falta de abordagens para a tarefa de ER, também é difícil apresentar resultados comparativos nesta fase, uma vez que ainda não há um conjunto de dados disponível para comparação.

Como resposta à pergunta de pesquisa definida no Capítulo 1, *"Como é possível abordar a tarefa de Extração de Relações para o domínio do Mercado Financeiro e idioma Português?"*, concluímos que há diversas formas para abordar a ER disponíveis na literatura, e que podem ser aplicadas no contexto desse trabalho. Porém, para que possamos desenvolver essa tarefa de ER, primeiramente foi necessário que construíssemos um conjunto de dados voltado para este escopo, visto que a literatura revisada aborda em sua maioria o idioma inglês. Após o desenvolvimento do corpus, construímos uma abordagem utilizando aprendizado de máquina em formato de pipeline, primeiro classificando as tuplas e posteriormente extraindo o trecho que sintetiza a relação semântica entre as ENs. Essa abordagem se mostrou mais adequada por possuir mais aderência ao contexto financeiro, no qual há muitas sentenças mas nem todas possuem uma relação semântica, e também por utilizar o *transformer* BERT, fazendo com que se reduza o esforço com extração de recursos ou mapeamento de regras.

Como principal contribuição deste trabalho, podemos citar o desenvolvimento de um modelo de Extração de Relação entre entidades nomeadas baseado em BERT, que substitui os recursos linguísticos explícitos, requeridos por outros métodos, como abordagens baseadas em regras ou que utilizam POS. Essa abordagem se torna muito mais simples, pois só precisa das informações da frase e do par de entidades concatenadas. Assim, permite o envio de mais de uma entrada, pois uma frase pode ter N pares de entidades nomeadas. Portanto, a abordagem adotada permite inferir que a sentença e o par de entidades são enviados separadamente. Outra contribuição deste trabalho encontra-se

no desenvolvimento de um grande corpus relacionado ao mercado financeiro, com textos anotados manualmente a partir de *tweets* e notícias fornecidas por analistas de IC para apoiar a tomada de decisão. Este corpus encontra-se disponível no GitHub [1] para que outros pesquisadores possam utilizá-lo para outras pesquisas.

Os resultados demonstram que a abordagem utilizada alcançou bons escores, atingindo uma pontuação de Jaccard de 76,3%, além de alcançar métricas gerais de *Recall* 87,1% e 84,5% de Precisão. Essa pontuação é interessante, pois o modelo foi capaz de extrair relações de diferentes tamanhos. Conforme mostra o Capítulo 6, o modelo foi mais penalizado ao extrair relações de apenas um *token*, geralmente formado por preposições.

Pelo fato de a abordagem utilizar o *transformer* BERT e não utilizar nenhum tipo de recurso que não seja da própria sentença, como regras de dependência, acreditamos que a método proposto pode ser aplicado em outros contextos diferentes do mercado financeiro. Contudo, para fazer este experimento, o conjunto de dados deste novo contexto deve estar anotado e formatado da mesma maneira que o disponibilizado neste estudo.

Mesmo que a construção do corpus e a sua disponibilização seja uma das contribuições deste estudo, é importante mencionar que o conjunto de dados não se caracteriza como uma Coleção Dourada. Ele foi anotado manualmente sem a correção de outros linguistas. Logo, como trabalho futuro, o corpus será incrementado com novos registros anotados e será transformado em uma Coleção Dourada, com o auxílio de linguistas para revisão das anotações e criação de medidas de concordância.

Também como sequência deste trabalho, será desenvolvido um terceiro modelo para inferir qual categoria de relação a tupla contém, como visto em outros trabalhos da literatura abordados no Capítulo 3. Para alcançar este objetivo, o corpus terá de ser anotado com mais uma coluna, que indique esta categoria a ser predita. Outro trabalho futuro consiste no desenvolvimento de uma Interface de Programação de Aplicação (API) para disponibilizar o modelo em uma página da web e, assim, permitir a utilização do mesmo para extração de relações pelo próprio usuário. Por fim, deseja-se desenvolver uma integração com banco de dados baseado em grafos, para que se possa persistir as relações extraídas neste banco de dados possibilitando a organização de uma base histórica de relações entre ENs do domínio financeiro.

[1] https://github.com/DanielReeyes/financial-market-corpus

REFERÊNCIAS BIBLIOGRÁFICAS

[Abr14] de Abreu, S. C. “Extração de relações do domínio de organizações para o português”, Tese de doutorado, Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS, Porto Alegre, Brazil, 2014, 112p.

[AJ18] Ameta, D.; Jat, P. M. “Information extraction from wikipedia articles using deepdive”. In: International Conference on Communication Information and Computing Technology, 2018, pp. 1–6.

[AT20] Alimova, I.; Tutubalina, E. “Multiple features for clinical relation extraction: a machine learning approach”, *Journal of Biomedical Informatics*, vol. 103, Mar 2020, pp. 1–9.

[BAT19] Bölücü, N.; Akgöl, D.; Tuç, S. “Bidirectional lstm-cnns with extended features for named entity recognition”. In: Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science, 2019, pp. 1–4.

[BB07] Bach, N.; Badaskar, S. “A survey on relation extraction”, *Language Technologies Institute, Carnegie Mellon University*, vol. 178, Nov 2007, pp. 15.

[BE08] Banko, M.; Etzioni, O. “The tradeoffs between open and traditional relation extraction”. In: Association for Computational Linguistics with the Human Language Technology Conference, 2008, pp. 28–36.

[BFS+13] Batista, D. S.; Forte, D.; Silva, R.; Martins, B.; Silva, M. “Extracçao de relaçoes semânticas de textos em português explorando a dbpédia e a wikipédia”, *Linguamatica*, vol. 5–1, Jul 2013, pp. 41–57.

[BMR+20] Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al.. “Language models are few-shot learners”, *ArXiv*, vol. abs/2005.14165, Mai 2020, pp. 1877–1901.

[BSF94] Bengio, Y.; Simard, P.; Frasconi, P. “Learning long-term dependencies with gradient descent is difficult”, *IEEE Transactions on Neural Networks*, vol. 5–2, Mar 1994, pp. 157–166.

[Car08] Cardoso, N. “Rembrandt-reconhecimento de entidades mencionadas baseado em relaçoes e análise detalhada do texto”, *Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM Linguateca*, vol. 1, Set 2008, pp. 195–211.

[CG19] Chen, J.; Gu, J. “Jointly extract entities and their relations from biomedical text”, *IEEE Access*, vol. 7, Nov 2019, pp. 162818–162827.

[CGC+20] Collovini, S.; Gonçalves, P. N.; Cavalheiro, G.; Santos, J.; Vieira, R. “Relation extraction for competitive intelligence”. In: International Conference on Computational Processing of the Portuguese Language, 2020, pp. 249–258.

[CGS+11] Collovini, S.; Grando, F.; Souza, M.; Freitas, L.; Vieira, R. “Semantic relations extraction in the organization domain”. In: International Conference on Applied Computing, 2011, pp. 99–106.

[CGSC20] Cabral, B. S.; Glauber, R.; Souza, M.; Claro, D. B. “Crossoie: Cross-lingual classifier for open information extraction.” In: International Conference on Computational Processing of Portuguese, 2020, pp. 368–378.

[Cha08] Chaves, M. “Geo-ontologias e padrões para reconhecimento de locais e de suas relações em textos: o sei-geo no segundo harem”, *Desafios na Avaliação Conjunta do Reconhecimento de Entidades Mencionadas: O Segundo HAREM Linguateca*, vol. 1, Set 2008, pp. 231–245.

[CK17] Chaniago, R.; Khodra, M. L. “Information extraction on novel text using machine learning and rule-based system”. In: International Conference on Innovative and Creative Information Technology, 2017, pp. 1–6.

[CLZZ19] Chen, D.; Li, B.; Zhou, C.; Zhu, X. “Automatically identifying bug entities and relations for bug analysis”. In: IEEE 1st International Workshop on Intelligent Bug Fixing, 2019, pp. 39–43.

[CMV16a] Collovini, S.; Machado, G.; Vieira, R. “Extracting and structuring open relations from portuguese text”. In: International Conference on Computational Processing of the Portuguese Language, 2016, pp. 153–164.

[CMV16b] Collovini, S.; Machado, G.; Vieira, R. “A sequence model approach to relation extraction in portuguese”. In: 10th International Conference on Language Resources and Evaluation, 2016, pp. 1908–1912.

[CPVV14] Collovini, S.; Pugens, L.; Vanin, A. A.; Vieira, R. “Extraction of relation descriptors for portuguese using conditional random fields”. In: Ibero-American Conference on Artificial Intelligence, 2014, pp. 108–119.

[CSC20] Cabral, B. S.; Souza, M.; Claro, D. B. “Explainable openie classifier with morpho-syntactic rules.” In: Hybrid Intelligence for Natural Language Processing Tasks, 2020, pp. 7–15.

[CT21] Christou, D.; Tsoumakas, G. “Improving distantly-supervised relation extraction through bert-based label and instance embeddings”, *IEEE Access*, vol. 9, Fev 2021, pp. 62574–62582.

[CUKR15] Chang, K.-W.; Upadhyay, S.; Kundu, G.; Roth, D. “Structural learning with amortized inference”. In: AAAI Conference on Artificial Intelligence, 2015, pp. 2525–2531.

[CV17] Collovini, S.; Vieira, R. “Relp: Portuguese open relation extraction”, *Knowledge Organization*, vol. 44–3, Jan 2017, pp. 163–177.

[CW08] Collobert, R.; Weston, J. “A unified architecture for natural language processing: Deep neural networks with multitask learning”. In: 25th International Conference on Machine Learning, 2008, pp. 160–167.

[CW18] Cruz, C. G. A.; Weitzel, L. “Evaluation of relation extraction systems for portuguese language pt-br”. In: 13th Iberian Conference on Information Systems and Technologies, 2018, pp. 1–6.

[CWB+11] Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. “Natural language processing (almost) from scratch”, *Journal of Machine Learning Research*, vol. 12, Nov 2011, pp. 2493–2537.

[CWY+20] Chen, Y.; Wang, K.; Yang, W.; Qing, Y.; Huang, R.; Chen, P. “A multi-channel deep neural network for relation extraction”, *IEEE Access*, vol. 8, Jan 2020, pp. 13195–13203.

[CX20] Cheng, W.; Xiong, J. “Entity relationship extraction based on bi-channel neural network”. In: 2nd International Conference on Machine Learning, Big Data and Business Intelligence, 2020, pp. 349–352.

[Dan13] Daniel, G. “Principles of artificial neural networks”. World Scientific, 2013, vol. 7, 384p.

[DCLT19] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171–4186.

[DJ16] Do, H. W.; Jeong, Y.-S. “Temporal relation classification with deep neural network”. In: International Conference on Big Data and Smart Computing, 2016, pp. 454–457.

[DLRBVM21] De Los Reyes, D.; Barcelos, A.; Vieira, R.; Manssour, I. “Related named entities classification in the economic-financial context”. In: European

Association for Computational Linguistics Hackashop on News Media Content Analysis and Automated Report Generation, 2021, pp. 8–15.

[DRR20] Deepa, C.; Raj, P. R.; Ramanujan, A. “Relation extraction across sentences using bi-directional long short term memory networks”. In: International Conference on Emerging Trends in Information Technology and Engineering, 2020, pp. 1–6.

[DSG14] Dos Santos, C.; Gatti, M. “Deep convolutional neural networks for sentiment analysis of short texts”. In: 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 69–78.

[dSN14] dos Santos Neto, J. F. “Reconhecimento de entidades nomeadas para o português usando redes neurais”, Dissertação de mestrado, Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS, Porto Alegre, Brazil, 2014, 95p.

[EFC+11] Etzioni, O.; Fader, A.; Christensen, J.; Soderland, S.; et al.. “Open information extraction: The second generation”. In: 22nd International Joint Conference on Artificial Intelligence, 2011, pp. 3–10.

[FGQ+17] Feng, X.; Guo, J.; Qin, B.; Liu, T.; Liu, Y. “Effective deep memory networks for distant supervised relation extraction.” In: International Joint Conference on Artificial Intelligence, 2017, pp. 4002–4008.

[FPPR19] Florez, E.; Precioso, F.; Pighetti, R.; Riveill, M. “Deep learning for identification of adverse drug reaction relations”. In: International Symposium on Signal Processing Systems, 2019, pp. 149–153.

[FPSS96] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. “From data mining to knowledge discovery in databases”, *AI Magazine*, vol. 17–3, Mar 1996, pp. 37.

[FSE11] Fader, A.; Soderland, S.; Etzioni, O. “Identifying relations for open information extraction”. In: Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1535–1545.

[GBC16] Goodfellow, I.; Bengio, Y.; Courville, A. “Deep learning”. MIT Press, 2016, 785p.

[GC18] Glauber, R.; Claro, D. B. “A systematic mapping study on open information extraction”, *Expert Systems with Applications*, vol. 112, Dez 2018, pp. 372–387.

[GCdO19] Glauber, R.; Claro, D. B.; de Oliveira, L. S. “Dependency parser on open information extraction for portuguese texts-dptoie and dependentie on iberlef.” In: Iberian Languages Evaluation Forum, 2019, pp. 442–448.

[GG11] Garcia, M.; Gamallo, P. “Dependency-based text compression for semantic relation extraction”. In: Recent Advances in Natural Language Processing Workshop on Information Extraction and Knowledge Acquisition, 2011, pp. 21–28.

[GGH19] Gan, T.; Gan, Y.; He, Y. “Subsequence-level entity attention lstm for relation extraction”. In: 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, 2019, pp. 262–265.

[GJM13] Graves, A.; Jaitly, N.; Mohamed, A.-r. “Hybrid speech recognition with deep bidirectional lstm”. In: IEEE workshop on Automatic Speech Recognition and Understanding, 2013, pp. 273–278.

[GM17] Gupta, S.; Manjhvar, A. K. “Relation classification from unstructured medical text using feature based machine learning approach”. In: International Conference on Trends in Electronics and Informatics, 2017, pp. 1135–1138.

[GMH13] Graves, A.; Mohamed, A.-r.; Hinton, G. “Speech recognition with deep recurrent neural networks”. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645–6649.

[Gra12] Graves, A. “Supervised sequence labelling”. In: *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer, 2012, cap. 2, pp. 5–13.

[Gru95] Gruber, T. R. “Toward principles for the design of ontologies used for knowledge sharing?”, *International Journal of Human-Computer Studies*, vol. 43–5-6, Nov 1995, pp. 907–928.

[GTM+21] Guimarães, D.; Trajano, D.; Manssour, I.; Vieira, R.; Bordini, R. H. “Entity relation extraction from news articles in portuguese for competitive intelligence based on bert”. In: Brazilian Conference on Intelligent System, 2021, pp. 1–15.

[GYJ+18] Gan, J.; Yang, J.; Jiao, D.; Li, S.; Chen, G. “Entity and relation extraction with dynamic memory cells”. In: International Conference on Network Infrastructure and Digital Content, 2018, pp. 11–15.

[HBF+01] Hochreiter, S.; Bengio, Y.; Frasconi, P.; Schmidhuber, J.; et al.. “Gradient flow in recurrent nets: the difficulty of learning long-term dependencies”. Capturado em: https://www.bioinf.jku.at/publications/older/ch7.pdf, Maio 2021.

[HKP11] Han, J.; Kamber, M.; Pei, J. “Data mining concepts and techniques third edition”, *The Morgan Kaufmann Series in Data Management Systems*, vol. 5–4, Jun 2011, pp. 83–124.

[HS97] Hochreiter, S.; Schmidhuber, J. “Long short-term memory”, *Neural Computation*, vol. 9–8, Nov 1997, pp. 1735–1780.

[HW20] Han, X.; Wang, L. “A novel document-level relation extraction method based on bert and entity information”, *IEEE Access*, vol. 8, Mai 2020, pp. 96912–96919.

[HZL20] He, J.; Zhao, Y.; Luo, G. “Dsrefc: Improving distantly-supervised neural relation extraction using feature combination”. In: 12th International Conference on Machine Learning and Computing, 2020, pp. 524–529.

[HZS19] Haihong, E.; Zhou, X.; Song, M. “Distant supervised relation extraction based on recurrent convolutional piecewise neural network”. In: International Symposium on Signal Processing Systems, 2019, pp. 169–175.

[HZSBHM15] Herrero-Zazo, M.; Segura-Bedmar, I.; Hastings, J.; Martínez, P. “Application of domain ontologies to natural language processing: a case study for drug-drug interactions”, *International Journal of Information Retrieval Research*, vol. 5–3, Jul 2015, pp. 19–38.

[Jac01] Jaccard, P. “Étude comparative de la distribution florale dans une portion des alpes et des jura”, *Bulletin de la Societe Vaudoise des Sciences Naturelles*, vol. 37, Mar 1901, pp. 547–579.

[JLHZ17] Ji, G.; Liu, K.; He, S.; Zhao, J. “Distant supervision for relation extraction with sentence-level attention and entity descriptions”. In: AAAI Conference on Artificial Intelligence, 2017, pp. 3060—3066.

[JM09] Jurafsky, D.; Martin, J. H. “Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”. Prentice Hall, 2009, 988p.

[Kim14] Kim, Y. “Convolutional neural networks for sentence classification”. In: Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1746–1751.

[Kit04] Kitchenham, B. “Procedures for performing systematic reviews”, *Keele University*, vol. 33, Jul 2004, pp. 1–26.

[KJFF15] Karpathy, A.; Johnson, J.; Fei-Fei, L. “Visualizing and understanding recurrent networks”, *ArXiv*, vol. abs/1506.02078, Jun 2015, pp. 11.

[KKvdSS96] Króse, B.; Krose, B.; van der Smagt, P.; Smagt, P. “An introduction to neural networks”, *Journal of Computer Science*, vol. 48, Nov 1996, pp. 123.

[KPGM20] Kumar, A.; Pandey, A.; Gadia, R.; Mishra, M. “Building knowledge graph using pre-trained language model for learning entity-aware relationships”. In: IEEE International Conference on Computing, Power and Communication Technologies, 2020, pp. 310–315.

[LBBH98] LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. “Gradient-based learning applied to document recognition”, *IEEE*, vol. 86–11, Nov 1998, pp. 2278–2324.

[LBH+15] Lu, J.; Behbood, V.; Hao, P.; Zuo, H.; Xue, S.; Zhang, G. “Transfer learning using computational intelligence: a survey”, *Knowledge-Based Systems*, vol. 80, Mai 2015, pp. 14–23.

[LCW16] Liang, S.; Chen, G.; Wang, W. “Learning mention and relation representation with convolutional neural networks for relation extraction”. In: IEEE International Conference on Network Infrastructure and Digital Content, 2016, pp. 437–441.

[LEF18] Lima, R.; Espinasse, B.; Freitas, F. “Ontoilper: an ontology-and inductive logic programming-based system to extract entities and relations from text”, *Knowledge and Information Systems*, vol. 56–1, Out 2018, pp. 223–255.

[LHCW19] Li, J.; Huang, G.; Chen, J.; Wang, Y. “Dual cnn for relation extraction with knowledge-based attention and word embeddings”, *Computational Intelligence and Neuroscience*, vol. 2019, Jan 2019, pp. 1–10.

[LJHZ16] Liu, H.; Jiang, C.; Hu, C.; Zhang, L. “Efficient relation extraction method based on spatial feature using elm”, *Neural Computing and Applications*, vol. 27–2, Dez 2016, pp. 271–281.

[LLS+21] Liu, H.; Li, Z.; Sheng, D.; Zheng, H.-T.; Shen, Y. “Multi-entity collaborative relation extraction”. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2021, pp. 7678–7682.

[LQP+19] Liu, B.; Qi, G.; Pan, L.; Duan, S.; Wu, T. “Incorporating human knowledge in neural relation extraction with reinforcement learning”. In: International Joint Conference on Neural Networks, 2019, pp. 1–8.

[LRW+18] Liu, J.; Ren, H.; Wu, M.; Wang, J.; Kim, H.-j. “Multiple relations extraction among multiple entities in unstructured text”, *Soft Computing*, vol. 22–13, Out 2018, pp. 4295–4305.

[LSC19] Lee, J.; Seo, S.; Choi, Y. S. “Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing”, *Symmetry*, vol. 11–6, Jan 2019, pp. 785.

[LVHM+16] Lossio-Ventura, J. A.; Hogan, W.; Modave, F.; Hicks, A.; Hanna, J.; Guo, Y.; He, Z.; Bian, J. “Towards an obesity-cancer knowledge base: Biomedical entity identification and relation detection”. In: IEEE International Conference on Bioinformatics and Biomedicine, 2016, pp. 1081–1088.

[LYC+20] Luo, L.; Yang, Z.; Cao, M.; Wang, L.; Zhang, Y.; Lin, H. “A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature”, *Journal of Biomedical Informatics*, vol. 103, Mar 2020, pp. 103384.

[LYL+18] Li, Q.; Yang, Z.; Luo, L.; Wang, L.; Zhang, Y.; Lin, H.; Wang, J.; Yang, L.; Xu, K.; Zhang, Y. “A multi-task learning based approach to biomedical entity relation extraction”. In: IEEE International Conference on Bioinformatics and Biomedicine, 2018, pp. 680–682.

[MA07] Morais, E. A. M.; Ambrósio, A. P. L. “Ontologias: conceitos, usos, tipos, metodologias, ferramentas e linguagens”, Relatório técnico, Instituto de Informática Universidade de Goiás, 2007, 21p.

[MMW10] Moncecchi, G.; Minel, J.-L.; Wonsever, D. “A survey of kernel methods for relation extraction”, *Workshop on NLP and Web-based Technologies*, vol. 1, 11 2010, pp. 1–9.

[MOM+20] Moreira, J.; Oliveira, C.; Macêdo, D.; Zanchettin, C.; Barbosa, L. “Distantly-supervised neural relation extraction with side information using bert”. In: International Joint Conference on Neural Networks, 2020, pp. 1–7.

[MSC+13] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; Dean, J. “Distributed representations of words and phrases and their compositionality”. In: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.

[MSO17] Mulang, I. O.; Singh, K.; Orlandi, F. “Matching natural language relations to knowledge graph properties for question answering”. In: 13th International Conference on Semantic Systems, 2017, pp. 89–96.

[NKD20] Nayak, A.; Kesri, V.; Dubey, R. K. “Knowledge graph based automated generation of test cases in software engineering”. In: 7th ACM International Conference on Data Science and Management of Data and 25th International Conference on Management of Data, 2020, pp. 289–295.

[NLW+19] Ni, J.; Liu, Y.; Wang, K.; Zhao, Z.; Sheng, Q. Z. “Distantly supervised relation extraction through a trade-off mechanism”. In: International Joint Conference on Neural Networks, 2019, pp. 1–8.

[NP17] Nayak, P.; Prajapati, G. L. “Extracting relation between brain region pairs from white text”. In: IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems, 2017, pp. 1–6.

[NTW17] Nguyen, D. B.; Theobald, M.; Weikum, G. “J-reed: joint relation extraction and entity disambiguation”. In: Conference on Information and Knowledge Management, 2017, pp. 2227–2230.

[PAA20] Parekh, A.; Anand, A.; Awekar, A. “Taxonomical hierarchy of canonicalized relations from multiple knowledge bases”. In: 7th ACM International Conference on Data Science and Management of Data and 25th International Conference on Management of Data, 2020, pp. 200–203.

[Pen13] Pena-Ayala, A. “Educational Data Mining: Applications and Trends”. Springer, 2013, vol. 524, 468p.

[Pen14] Pena-Ayala, A. “Educational data mining: A survey and a data mining-based analysis of recent works”, *Expert Systems with Applications*, vol. 41–4, Mar 2014, pp. 1432–1462.

[PHT+19] Peng, M.; Hu, W.; Tian, G.; Wang, B.; Wang, H.; Wang, G. “Dilated convolutional networks incorporating soft entity type constraints for distant supervised relation extraction”. In: International Joint Conference on Neural Networks, 2019, pp. 1–7.

[PIW+17] Pandey, C.; Ibrahim, Z.; Wu, H.; Iqbal, E.; Dobson, R. “Improving rnn with attention and embedding for adverse drug reactions”. In: International Conference on Digital Health, 2017, pp. 67–71.

[PLL+19] Pang, Y.; Liu, J.; Liu, L.; Yu, Z.; Zhang, K. “A deep neural network model for joint entity and relation extraction”, *IEEE Access*, vol. 7, Out 2019, pp. 179143–179150.

[PPM+19] Pingle, A.; Piplai, A.; Mittal, S.; Joshi, A.; Holt, J.; Zak, R. “Relext: Relation extraction using deep learning approaches for cybersecurity knowledge graph improvement”. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019, pp. 879–886.

[PST+19] Phi, V.-T.; Santoso, J.; Tran, V.-H.; Shindo, H.; Shimbo, M.; Matsumoto, Y. “Distant supervision for relation extraction via piecewise attention and bag-level contextual inference”, *IEEE Access*, vol. 7, Jul 2019, pp. 103570–103582.

[PY10] Pan, S. J.; Yang, Q. “A survey on transfer learning”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, Out 2010, pp. 1345–1359.

[PYCX18] Pan, Q.; Yu, C.; Chen, D.; Xiang, L. “Joint extraction of entities and relations of breast ultrasound reports based on deep learning”. In: IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems, 2018, pp. 219–225.

[QWR14] Quan, C.; Wang, M.; Ren, F. “An unsupervised text mining method for relation extraction from biomedical literature”, *PloS One*, vol. 9–7, Jul 2014, pp. e102039.

[QXG17] Qin, P.; Xu, W.; Guo, J. “Designing an adaptive attention mechanism for relation classification”. In: International Joint Conference on Neural Networks, 2017, pp. 4356–4362.

[QZZ20] Qin, F.; Zhang, Z.; Zheng, X. “A joint course knowledge entity and relation extraction method for educational data”. In: International Conference on Information Science and Education, 2020, pp. 570–576.

[RFJ19] Ren, Y.; Fei, H.; Ji, D. “Drug-drug interaction extraction using a span-based neural network model”. In: IEEE International Conference on Bioinformatics and Biomedicine, 2019, pp. 1237–1239.

[RHW86] Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. “Learning representations by back-propagating errors”, *Nature*, vol. 323–6088, Out 1986, pp. 533–536.

[RNSS18] Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. “Improving language understanding by generative pre-training”, *ArXiv*, vol. abs/2110.05448, Jun 2018, pp. 1–12.

[RWH+17] Ren, X.; Wu, Z.; He, W.; Qu, M.; Voss, C. R.; Ji, H.; Abdelzaher, T. F.; Han, J. “Cotype: Joint extraction of typed entities and relations with knowledge bases”. In: 26th International Conference on World Wide Web, 2017, pp. 1015–1024.

[RWP15] Romadhony, A.; Widyantoro, D. H.; Purwarianti, A. “Phrase-based clause extraction for open information extraction system”. In: International Conference on Advanced Computer Science and Information Systems, 2015, pp. 155–162.

[Sar08] Sarawagi, S. “Information Extraction”. Now Publishers Inc, 2008, 116p.

[SC20] Sena, C. F. L.; Claro, D. B. “Pragmaticoie: a pragmatic open information extraction for portuguese language”, *Knowledge and Information Systems*, vol. 62, Set 2020, pp. 1–26.

[SdSK+20] Schneider, E. T. R.; de Souza, J. V. A.; Knafou, J.; e Oliveira, L. E. S.; Copara, J.; Gumiel, Y. B.; de Oliveira, L. F. A.; Paraiso, E. C.; Teodoro, D.; Barra, C. M. C. M. “Biobertpt-a portuguese neural language model for clinical named entity recognition”. In: 3rd Clinical Natural Language Processing Workshop, 2020, pp. 65–72.

[SG17] Shi, W.; Gao, S. “Relation extraction via position-enhanced convolutional neural network”. In: International Conference on Intelligent Environments, 2017, pp. 142–148.

[SHL19] Shi, M.; Huang, J.; Li, C. “Entity relationship extraction based on blstm model”. In: IEEE/ACIS 18th International Conference on Computer and Information Science, 2019, pp. 266–269.

[SJC+18] Su, S.; Jia, N.; Cheng, X.; Zhu, S.; Li, R. “Exploring encoder-decoder model for distant supervised relation extraction.” In: International Joint Conference on Artificial Intelligence, 2018, pp. 4389–4395.

[SKL+15] Song, M.; Kim, W. C.; Lee, D.; Heo, G. E.; Kang, K. Y. “Pkde4j: Entity and relation extraction for public knowledge discovery”, *Journal of Biomedical Informatics*, vol. 57, Out 2015, pp. 320–332.

[SNL20] Souza, F.; Nogueira, R.; Lotufo, R. “BERTimbau: pretrained BERT models for Brazilian Portuguese”. In: 9th Brazilian Conference on Intelligent Systems, 2020, pp. 1–8.

[SSJ+18] Shen, Y.; Sun, J.; Jia, P.; Zhang, L.; Han, D.; Shen, X.; Li, Y. “Entity-dependent long-short time memory network for semantic relation extraction”. In: 5th IEEE International Conference on Cloud Computing and Intelligence Systems, 2018, pp. 762–766.

[SVL14] Sutskever, I.; Vinyals, O.; Le, Q. V. “Sequence to sequence learning with neural networks”. In: Advances in Neural Information Processing Systems, 2014, pp. 3104–3112.

[T+06] Tan, P.-N.; et al.. “Introduction to data mining”. Pearson Education India, 2006, 864p.

[Tan19] Tang, Y. “An extended sequence labeling approach for relation extraction”. In: IEEE International Conference on Power, Intelligent Computing and Systems, 2019, pp. 121–124.

[TC14] Taba, L. S.; Caseli, H. “Automatic semantic relation extraction from portuguese texts”. In: 9th International Conference on Language Resources and Evaluation, 2014, pp. 2739–2746.

[TGQ17] Tai, L.; Guo, F.; Qin, S. “Semi-supervised entity relation extraction based on trigger word”. In: 3rd IEEE International Conference on Computer and Communications, 2017, pp. 497–501.

[TKK21] Tran, T.; Kavuluru, R.; Kilicoglu, H. “Attention-gated graph convolutions for extracting drug interaction information from drug labels”, *ACM Transactions on Computing for Healthcare*, vol. 2–2, Mar 2021, pp. 1–19.

[TQG17] Tai, L.; Qin, S.; Guo, F. “A pattern learning method based on kernel function”. In: 2nd International Conference on Communication and Information Systems, 2017, pp. 324–328.

[TSM15] Tai, K. S.; Socher, R.; Manning, C. D. “Improved semantic representations from tree-structured long short-term memory networks”. In: 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015, pp. 1556–1566.

[VSP+17] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. “Attention is all you need”. In: Advances in neural information processing systems, 2017, pp. 5998–6008.

[WC20] Wu, C.; Chen, L. “Utber: Utilizing fine-grained entity types to relation extraction with distant supervision”. In: IEEE International Conference on Smart Data Services, 2020, pp. 63–71.

[Wei02] Weiss, A. “A brief guide to competitive intelligence: how to gather and use information on competitors”, *Business Information Review*, vol. 19–2, Jun 2002, pp. 39–47.

[WH19] Wu, S.; He, Y. “Enriching pre-trained language model with entity information for relation classification”. In: 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 2361–2364.

[WLR+19] Wang, C.; Li, Z.; Ren, S.; Wang, H.; Hu, F.; Deng, W. “Extraction method with word distribution enriched deep residual network”. In: 3rd International Conference on Advances in Image Processing, 2019, pp. 211–218.

[WLYL20] Wang, Q.; Lv, L.; Yu, B.; Li, S. “End-to-end relation extraction using graph convolutional network with a novel entity attention”. In: IEEE 6th International Conference on Computer and Communications, 2020, pp. 2086–2093.

[WLZL20] Wu, H.; Lei, Q.; Zhang, X.; Luo, Z. “Creating a large-scale financial news corpus for relation extraction”. In: 3rd International Conference on Artificial Intelligence and Big Data, 2020, pp. 259–263.

[WSM+20] Wang, Y.; Sun, Y.; Ma, Z.; Gao, L.; Xu, Y.; Wu, Y. “A method of relation extraction using pre-training models”. In: 13th International Symposium on Computational Intelligence and Design, 2020, pp. 176–179.

[WWMW20] Wang, C.; Wang, Y.; Mo, J.; Wang, S. “End-to-end relation extraction based on part of speech syntax tree”. In: 2nd International Conference on Machine Learning, Big Data and Business Intelligence, 2020, pp. 5–9.

[WXD20] Wang, L.; Xiong, C.; Deng, N. “A research on overlapping relationship extraction based on multi-objective dependency”. In: 15th International Conference on Computer Science & Education, 2020, pp. 618–622.

[WXG19] Wang, Y.; Xin, X.; Guo, P. “Relation extraction via attention-based cnns using token-level representations”. In: 15th International Conference on Computational Intelligence and Security, 2019, pp. 113–117.

[WZCL18] Wang, S.; Zhang, Y.; Che, W.; Liu, T. “Joint extraction of entities and relations based on a novel graph scheme.” In: 27th International Joint Conference on Artificial Intelligence, 2018, pp. 4461–4467.

[XDFX18] Xu, T.; Du, Y.; Fu, C.; Xie, C. “Incorporating forward and backward instances in a bi-lstm-cnn model for relation classification”. In: IEEE 4th International Conference on Computer and Communications, 2018, pp. 2133–2137.

[XdLS15] Xavier, C. C.; de Lima, V. L. S.; Souza, M. “Open information extraction based on lexical semantics”, *Journal of the Brazilian Computer Society*, vol. 21–1, Mai 2015, pp. 1–14.

[XQP18] Xue, L.; Qing, S.; Pengzhou, Z. “Relation extraction based on deep learning”. In: IEEE/ACIS 17th International Conference on Computer and Information Science, 2018, pp. 687–691.

[XZM+19] Xue, K.; Zhou, Y.; Ma, Z.; Ruan, T.; Zhang, H.; He, P. “Fine-tuning bert for joint entity and relation extraction in chinese medical text”. In: IEEE International Conference on Bioinformatics and Biomedicine, 2019, pp. 892–897.

[YCC+20] Yu, H.; Cao, Y.; Cheng, G.; Xie, P.; Yang, Y.; Yu, P. “Relation extraction with bert-based pre-trained model”. In: International Wireless Communications and Mobile Computing, 2020, pp. 1382–1387.

[YDHX20] Yin, J.; Duan, P.; Huang, W.; Xiong, S. “Probabilistic graph attention for relation extraction for domain of geography”. In: 3rd International Conference on Algorithms, Computing and Artificial Intelligence, 2020, pp. 1–6.

[YDZ+19] Yao, H.; Dong, L.; Zhen, S.; Kang, X.; Li, X.; Liang, Q. “Distant-supervised relation extraction with hierarchical attention based on knowledge graph”. In: IEEE 31st International Conference on Tools with Artificial Intelligence, 2019, pp. 229–236.

[YGJ+17] Yuan, J.; Guo, H.; Jin, Z.; Jin, H.; Zhang, X.; Luo, J. “One-shot learning for fine-grained relation extraction via convolutional siamese neural network”. In: IEEE International Conference on Big Data, 2017, pp. 2194–2199.

[YH19] Yi, R.; Hu, W. “Pre-trained bert-gru model for relation extraction”. In: 8th International Conference on Computing and Pattern Recognition, 2019, pp. 453–457.

[YJTC20] Yu, E.; Jia, Y.; Tian, Y.; Chang, Y. “A two-level noise-tolerant model for relation extraction with reinforcement learning”. In: IEEE International Conference on Knowledge Graph, 2020, pp. 367–373.

[YJW+20] Yu, E.; Jia, Y.; Wang, S.; Li, F.; Chang, Y. “Context and type enhanced representation learning for relation extraction”. In: IEEE International Conference on Knowledge Graph, 2020, pp. 329–335.

[YSW20] Yin, B.; Sun, Y.; Wang, Y. “Entity relation extraction method based on fusion of multiple information and attention mechanism”. In: IEEE 6th International Conference on Computer and Communications, 2020, pp. 2485–2490.

[YZZ20] Yang, L.; Zheng, L.; Zheng, L. “Research on extraction of human information entity relationship based on improved capsule network”. In: International Workshop on Electronic Communication and Artificial Intelligence, 2020, pp. 41–45.

[ŽB15] Žitnik, S.; Bajec, M. “Iterative joint extraction of entities, relationships and coreferences from text sources”. In: IEEE 9th International Conference on Research Challenges in Information Science, 2015, pp. 412–422.

[ZCL17] Zhang, Q.; Chen, M.; Liu, L. “A review on entity relation extraction”. In: Second International Conference on Mechanical, Control and Computer Engineering, 2017, pp. 178–183.

[ZLL+19] Zhou, X.; Liu, L.; Luo, X.; Chen, H.; Qing, L.; He, X. “Joint entity and relation extraction based on reinforcement learning”, *IEEE Access*, vol. 7, Set 2019, pp. 1–12.

[ZLW+16] Zheng, H.; Li, Z.; Wang, S.; Yan, Z.; Zhou, J. “Aggregating inter-sentence information to enhance relation extraction”. In: 30th Association for the

Advancement of Artificial Intelligence Conference on Artificial Intelligence, 2016, pp. 3108–3114.

[ZLWX19] Zhou, K.; Luo, X.; Wang, H.; Xu, R. “Multi-task learning for relation extraction”. In: IEEE 31st International Conference on Tools with Artificial Intelligence, 2019, pp. 1480–1487.

[ZLWZ17] Zhang, Q.; Liu, J.; Wang, Y.; Zhang, Z. “A convolutional neural network method for relation classification”. In: International Conference on Progress in Informatics and Computing, 2017, pp. 440–444.

[ZSZ19] Zhu, Z.; Su, J.; Zhou, Y. “Improving distantly supervised relation classification with attention and semantic weight”, *IEEE Access*, vol. 7, Jun 2019, pp. 91160–91168.

[ZW17] Zhang, Y.; Wallace, B. C. “A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification”. In: The 8th International Joint Conference on Natural Language Processin, 2017, pp. 3111–3119.

[ZYS+17] Zhao, Z.; Yang, Z.; Sun, C.; Wang, L.; Lin, H. “A hybrid protein-protein interaction triple extraction method for biomedical literature”. In: IEEE International Conference on Bioinformatics and Biomedicine, 2017, pp. 1515–1521.

[ZZ18] Zhou, Z.; Zhang, H. “Research on entity relationship extraction in financial and economic field based on deep learning”. In: IEEE 4th International Conference on Computer and Communications, 2018, pp. 2430–2435.

[ZZ20] Zhong, L.; Zhu, Y. “Relation extraction with proactive domain adaptation strategy”. In: IEEE International Conference on Knowledge Graph, 2020, pp. 441–448.

[ZZX+15] Zhang, C.; Zhang, Y.; Xu, W.; Ma, Z.; Leng, Y.; Guo, J. “Mining activation force defined dependency patterns for relation extraction”, *Knowledge-Based Systems*, vol. 86, Set 2015, pp. 278–287.

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Graduação
Av. Ipiranga, 6681 - Prédio 1 - 3º. andar
Porto Alegre - RS - Brasil
Fone: (51) 3320-3500 - Fax: (51) 3339-1564
E-mail: prograd@pucrs.br
Site: www.pucrs.br