Reforço e Delimitação Contextual para Reconhecimento de Entidades e Relações em Documentos Oficiais

Fabiano Muniz Belém, Marcelo Ganem, Celso França, Marcos Carvalho, Alberto H. F. Laender, Marcos André Gonçalves

[1] Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

{fmuniz,masganem,celsofranca,marcoscarvalho,laender,mgoncalv}@dcc.ufmg.br

Abstract. *Transformer architectures have become the main component of various state-of-the-art methods for natural language processing tasks, such as Named Entity Recognition and Relation Extraction (NER+RE). As these architectures rely on semantic aspects of word sequences, they may fail to accurately identify and delimit entity spans when there is little semantic context surrounding the named entities. This is the case of entities composed by digits and punctuation only, such as IDs and phone numbers, as well as long composed names. In this paper, we propose new techniques for contextual reinforcement and entity delimitation based on pre- and post-processing techniques to provide a richer semantic context, improving SpERT, a state-of-the-art Span-based Entity and Relation Transformer. We evaluate our strategies using real data from public administration documents and court lawsuits. Our results show that our pre- and post-processing strategies, when used co-jointly, allows significant improvements on NER+ER effectiveness.*

Resumo. *Arquiteturas neurais baseadas em transformers tornaram-se o principal componente de vários métodos do estado-da-arte em tarefas de processamento de linguagem natural, tais como Reconhecimento de Entidades Nomeadas e Extração de Relações (REN+ER). Como essas arquiteturas baseiam-se em aspectos semânticos de sequências de palavras, elas podem não funcionar na identificação e delimitação de entidades nomeadas quando há pouco contexto semântico associado, tais como entidades compostas por dígitos e pontuações apenas (e.g., números de CPF) e entidades com nomes compostos. Neste artigo, são propostas novas técnicas de reforço contextual e delimitação de entidades baseadas em pré- e pós-processamento de dados para enriquecer o contexto semântico, melhorando assim um método do estado-da-arte para REN+RE, o SpERT (Span-Based Entity and Relation Transformer). Tais técnicas foram avaliadas usando dados reais de diários oficiais e de processos judiciais. Os resultados mostram que, quando aplicadas em conjunto, as estratégias de pré- e pós-processamento levam a ganhos significativos na efetividade de REN+ER.*

1. Introdução

As tarefas de Reconhecimento de Entidades Nomeadas (REN) e Extração de Relações (ER) são úteis em diversas aplicações de gerência de dados como deduplicação de registros [Silva et al. 2019], integração de dados [Brunner & Stockinger 2020], construção

de bases de conhecimento [Niu et al. 2012] e busca em coleções de textos não-estruturados [Caputo et al. 2009]. No contexto dessas aplicações, enquanto a tarefa de REN visa identificar as entidades mencionadas em um texto (e.g., nomes de pessoas e de organizações), bem como classificá-las em um conjunto pré-definido de categorias, a tarefa de ER procura identificar e classificar as possíveis relações existentes entre as entidades do texto (e.g., relação *empregado-empregador* existente entre uma pessoa e uma organização) [Eberts & Ulges 2020].

Nesse contexto, arquiteturas neurais baseadas em *transformers* (e.g., *Bidirectional Encoders from Transformers* (BERT) [Devlin et al. 2019]) constituem o estado-da-arte em REN e ER. Dentre esses métodos, destaca-se o *Span-based Entity and Relation Transformer* (SpERT) [Eberts & Ulges 2020], que realiza conjuntamente as duas tarefas mencionadas acima. O SpERT, assim como outros métodos que utilizam *transformers*, baseiam-se em aspectos semânticos (i.e., contextuais) de sequências de palavras, que podem estar ausentes, por exemplo, em entidades compostas por dígitos e pontuações apenas (e.g., números de identidade e de telefone). Ao mesmo tempo, esses tipos de entidade apresentam padrões bem comportados que podem ser capturados por expressões regulares, as quais podem ser utilizadas para realçar e reforçar o contexto semântico das frases onde tais padrões ocorrem.

Assim, o principal objetivo deste artigo é propor novas técnicas que realcem o contexto semântico utilizado para fazer o reconhecimento de entidades e relações. Para isso, foi proposta uma nova etapa de pré-processamento do texto de entrada. Essa etapa utiliza expressões regulares para destacar no texto entidades como números de CPF (Cadastro de Pessoa Física) e de CNPJ (Cadastro Nacional de Pessoa Jurídica), o que ajuda a reconhecer não apenas estes tipos de entidades mais regulares, como também entidades que ocorrem próximas a eles ou no mesmo contexto – por exemplo, um CNPJ frequentemente ocorre próximo ao nome da organização a ele associada.

Apesar das melhorias potenciais dessa técnica de reforço semântico, foi verificado empiricamente que o SpERT original pode retornar diferentes partes de uma única entidade como se fossem entidades diferentes, e.g., "João da Silva" e apenas "João" quando nome e sobrenome aparecem juntos. Ele também pode incluir palavras que não fazem parte da entidade (e.g., em "SECRETÁRIO DE ESTADO REGISTRA AFASTAMENTO", o método erroneamente incluiu a expressão "REGISTRA AFASTAMENTO" como parte da entidade, provavelmente devido ao padrão de maiúsculas do texto). Para melhor delimitar as entidades, foi proposta também uma etapa de pós-processamento que visa unificar entidades que não foram bem delimitadas pelo SpERT e métodos similares. Esta etapa escolhe a menção mais provável para cada conjunto de menções que estejam sobrepostas na saída do método original. A probabilidade é dada pela própria saída do SpERT, que associa um *score* a cada entidade reconhecida.

As novas técnicas foram avaliadas utilizando dois conjuntos de dados: um de diários oficiais e outro de processos judiciais. Os dados de processos judiciais foram obtidos da coleção LENER-BR [Luz de Araujo et al. 2018], enquanto que os dados de diários oficiais consistem em dados previamente coletados do Diário Oficial do Estado de Minas Gerais (DOMG), compreendendo todos os 208 documentos do diário oficial do ano de 2016. Utilizando um conjunto predefinido de 10 rótulos de entidades e 10 rótulos de relações, foi rotulada manualmente uma amostra representativa dos dados do DOMG,

selecionada através de técnicas de Aprendizado Ativo [Wang et al. 2021]. Para ampliar o conjunto de dados de treino nessa coleção, melhorando o treinamento do método SpERT, foram utilizadas técnicas de geração de dados de treino sintéticos, através da formulação de frases que relacionam as entidades previamente identificadas.

Os resultados experimentais mostram que a estratégia de pré-processamento leva a ganhos de 4% em revocação nas tarefas de REN e ER, enquanto a etapa de pós-processamento é responsável por ganhos de até 17% em precisão na tarefa de REN e até 32% na tarefa de ER, permitindo delimitar de forma mais precisa as menções a entidades no texto. Ambas estratégias propostas apresentam um custo adicional desprezível (inferior a 2%) em relação ao custo total do reconhecimento de entidades e relações.

Em suma, as principais contribuições deste artigo são:

- Uma técnica de realce de contexto semântico baseada em pré-processamento dos dados de entrada, gerando resultados com maior cobertura de entidades e relações (revocação);
- Uma técnica de delimitação de entidades no texto por meio de pós-processamento dos resultados, aumentando a precisão no reconhecimento tanto de entidades como de relações;
- Geração de uma nova coleção de dados relevantes para avaliação e treino de algoritmos de REN+ER em documentos oficiais, disponibilizados publicamente[1].

Vale ressaltar que as estratégias propostas neste artigo estão sendo utilizadas por aplicações finalísticas de apoio ao Ministério Público de Minas Gerais. Dentre elas, podemos citar uma ferramenta de busca em documentos oficiais e uma ferramenta de classificação semântica.

O restante deste artigo está organizado da seguinte maneira. A Seção 2 apresenta trabalhos relacionados, enquanto a Seção 3 apresenta a definição do problema abordado. A Seção 4 descreve as estratégias propostas, enquanto a Seção 5 descreve a metodologia de avaliação, cujos resultados experimentais são apresentados e discutidos na Seção 6. Por fim, a Seção 7 conclui o artigo e aponta algumas direções para trabalhos futuros.

2. Trabalhos Relacionados

Estratégias de REN e ER podem ser divididas em dois tipos de abordagem: *token-based* e *span-based*. A primeira classifica cada palavra ou *token* do texto em um dos tipos de entidade, adicionalmente identificando se a palavra pertence ao início, meio ou fim da entidade identificada [Finkel et al. 2005, Patil et al. 2020]. Já as estratégias *span-based* primeiro enumeram todos os *spans* (sequências de *tokens*) que sejam menores que um dado limite e, em seguida, classificam cada um dos *spans* enumerados [Eberts & Ulges 2020, Fu et al. 2021, Liu et al. 2021]. Adicionalmente, para cada par de *spans* com alta probabilidade de se tratar de uma entidade, tais técnicas inferem se há uma relação entre elas e qual o tipo de relação.

Um exemplo de método *token-based* tradicionalmente utilizado em REN são os baseados em *Conditional Random Fields* (CRFs) [Finkel et al. 2005, Patil et al. 2020]. Os CRFs são modelos probabilísticos que inferem a categoria de cada token t de um texto

[1]https://www.kaggle.com/datasets/fabianomunizbelem/domg-labeled-entities-and-relations

explorando atributos de t (e.g., padrão de letras maiúsculas e minúsculas) e atributos dos *tokens* adjacentes a t, bem como as inferências realizadas para os mesmos.

Mais recentemente, estratégias *span-based* [Fu et al. 2021] têm recebido maior atenção por produzirem resultados superiores às abordagens *token-based* e serem facilmente modeladas através de arquiteturas neurais baseadas em *transformers*, utilizando apenas atributos "crus" (*word embeddings*), sem a necessidade de elaborar e extrair atributos complexos dos dados. Além disso, tais estratégias permitem a extração de entidades que se sobrepõem no texto. Por exemplo, no caso de "Ministério Público de Minas Gerais", poderiam ser consideradas duas entidades: "Ministério Público de Minas Gerais" e "Minas Gerais". Entretanto, uma desvantagem dessa estratégia, como mostram nossos experimentos, é que ela é frequentemente imprecisa na delimitação de uma entidade no texto. Para tratar tal problema, é proposto neste artigo a incorporação às técnicas do estado-da-arte de uma estratégia de pós-processamento de resultados de REN+ER que torna a delimitação das entidades mais precisa, além de não depender de treino adicional.

Redes neurais baseadas em *transformers* (e.g., *Bidirectional Encoders from Transformers* ou BERT [Devlin et al. 2019]) representam o estado-da-arte em diversas tarefas de processamento de linguagem natural, inclusive na tarefa de reconhecimento de entidades e suas relações (REN+ER). Dentre os métodos de REN+ER baseados em transformers, destaca-se o *Span-based Entity and Relation Transformer* (SpERT) [Eberts & Ulges 2020], que codifica *spans* do texto em uma representação vetorial (*embeddings*) baseada em modelos pré-treinados para classificá-los em uma das categorias pré-definidas de entidades ou eventualmente como uma "não-entidade". Pares de *spans* que o algoritmo identifica como entidades são também representados no espaço vetorial e eventualmente atribuídos a alguma categoria pré-definida de relação. Eberts & Ulges [2021] estendem a arquitetura do SpERT para incluir um agrupamento de menções que se referem a uma mesma entidade em diferentes segmentos de um texto.

Entretanto, métodos como o SpERT são fortemente baseados no contexto semântico de sequências de palavras no texto, podendo apresentar dificuldades no reconhecimento de entidades e relações quando há pouca informação semântica nessas sequências de palavras (e.g., entidades formadas apenas por dígitos e pontuações, como números de CPF e de telefone). Neste artigo, este problema é contornado por meio de um reforço contextual dado pela marcação dessas entidades por meio de expressões regulares, que realçam o contexto semântico do texto. Essa nova etapa auxilia na identificação não apenas de entidades mais regulares, como também das entidades próximas a elas. Como consequência, há também melhoria no reconhecimento de relações entre as entidades.

3. Descrição das Tarefas

Nesta seção, são descritas as tarefas de Reconhecimento de Entidades Nomeadas (REN) e de Extração de Relações (ER). A primeira busca extrair e classificar entidades mencionadas em textos, possibilitando a sua separação em categorias pré-definidas tais como: nomes de pessoas, locais e organizações, valores monetários, CPFs, CNPJs, números de telefone, entre outras. Tipicamente, a tarefa REN processa um texto não estruturado, como o do seguinte exemplo fictício "A empresa XYZ Ltda., inscrita no CNPJ 12.345.678/1234-56, está localizada na Rua A, 1001", produzindo um bloco de texto anotado que destaca as entidades nomeadas e suas respectivas categorias, com as marcações

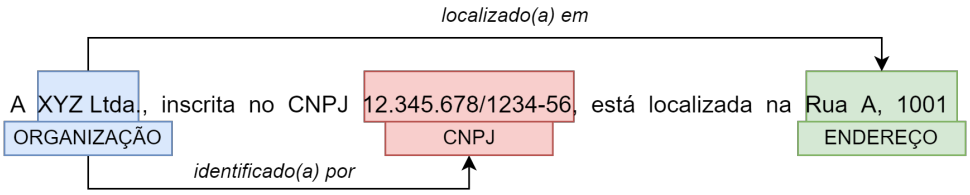

Figura 1. Exemplo de uma tarefa de REN+ER.

"ORGANIZAÇÃO", "CNPJ" e "ENDEREÇO" conforme mostrado na Figura 1.

Já a tarefa de ER retorna como resultado uma lista de relações entre as entidades encontradas no texto, representadas como setas na Figura 1. Formalmente, uma relação pode ser definida como uma tripla (e_1, r, e_2), onde e_1 e e_2 são entidades e r é o tipo da relação entre elas. Neste artigo, é tratado o problema do reconhecimento conjunto de entidades e relações, abreviado como REN+ER.

4. Estratégias Propostas

Foi proposta a extensão de estratégias de REN+ER através de uma técnica de pré-processamento do texto de entrada aqui denominada *Reforço Contextual* (ReCon) e uma estratégia de pós-processamento do resultado, a *Delimitação de Entidades* (DEnt). A técnica de ReCon (Subseção 4.1) pode ser aplicada a qualquer conjunto de dados que tenha pelo menos um tipo de entidade regular, i.e., que pode ser identificado por meio de expressões regulares. Já a estratégia DEnt (Subseção 4.2) pode ser aplicada a qualquer método que produza uma estimativa da confiança na classificação das entidades. Como esses cenários são bastante comuns em REN+ER [Zhang et al. 2018], tais estratégias são largamente aplicáveis. Neste artigo foi utilizado como estratégia-base o *Span-based Entity and Relation Transformer* (SpERT) [Eberts & Ulges 2020], por ele fazer parte do estado-da-arte em REN+ER.

4.1. Reforço Contextual (ReCon)

A estratégia de ReCon consiste em inserir, no texto, marcações preliminares que indiquem o tipo, o início e o fim de algumas das entidades listadas no texto. Para que isso seja possível, foca-se em entidades que possuem um padrão mais "bem comportado" que possa ser facilmente capturado por expressões regulares.

No caso particular da coleção de diários oficiais, foram identificados diversos tipos de entidade com características regulares. Um deles é o CPF, que é formado por 11 dígitos, eventualmente intercalados por pontos e por um hífen que separa os dois dígitos verificadores dos demais dígitos. Isso pode ser descrito pela seguinte expressão regular: $\backslash d\{3\}\backslash .?\backslash d\{3\}\backslash .?\backslash d\{3\}\backslash -?\backslash d\{2\}$. Essa e outras expressões regulares úteis para a identificação de entidades em documentos oficiais podem ser encontradas na página do repositório (vide *link* na Seção 1).

Dessa forma, para cada sequência de caracteres do texto que casar com uma das expressões regulares, insere-se a expressão [*tipo*], imediatamente antes e depois da sequência identificada, onde *tipo* é um nome correspondente ao tipo de entidade capturado pela expressão regular (e.g., CPF, DATA). São marcados tanto os dados que são

fornecidos como treino à estratégia-base, como os novos dados para os quais se deseja extrair entidades e relações. Após a extração, é possível remover as marcações para que o resultado seja exibido.

Uma alternativa para as marcações seria fazer o reconhecimento das entidades regulares diretamente com expressões regulares, sem passar pela estratégia-base. No entanto, foi verificado que as marcações também ajudam tal método na identificação de outras entidades próximas às regulares.

4.2. Delimitação de Entidades (DEnt)

Um outro desafio na tarefa REN+ER é a frequente dificuldade em delimitar algumas menções a entidades. Por exemplo, o algoritmo pode não capturar o nome completo de uma pessoa caso ele inclua muitos sobrenomes. Tratar isso baseando-se na variação do padrão de letras maiúsculas/minúsculas pode não funcionar em documentos oficiais, que muitas vezes apresentam o restante da frase em caixa alta, dificultando a distinção entre nomes próprios e palavras comuns. O mesmo vale para nomes de organizações ou até mesmo para entidades regulares, dependendo dos padrões de espaçamento e pontuações entre dígitos (e.g., "14/06" VS. "14 / 06" VS. "14-06", etc).

Nesses casos, as abordagens *span-based* retornam diferentes sequências que se referem a partes de uma mesma entidade (ou mesmo inclui palavras que estão fora dos limites de uma entidade). Isso garante a manutenção da revocação das menções, mas danifica a precisão no reconhecimento de entidades e consequentemente de relações.

Assim, foi proposta uma estratégia de Delimitação de Entidades (DEnt) que unifica as menções que se referem à mesma entidade utilizando as estimativas de confiança na classificação da própria estratégia-base (no caso, o SpERT). Para isso, para cada conjunto de menções a entidades que se sobrepõem na saída do algoritmo, escolhemos apenas a menção mais provável, de acordo com a estimativa do algoritmo.

A Figura 2 ilustra as etapas propostas para a tarefa REN+ER, apresentando um exemplo e os resultados esperados para o mesmo. Na etapa (a), o ReCon identificou uma sequência de caracteres que casa com a expressão regular para CNPJs, marcando essa sequência, o que facilitou a posterior identificação dessa entidade pelo método-base na etapa (b). O resultado do método-base encontrou duas entidades sobrepostas: "XYZ" e "XYZ Ltda.", com 85% e 90% de chance, respectivamente, de se tratarem de uma ORGANIZAÇÃO. Na etapa final (c), a DEnt elimina essa "duplicidade", mantendo apenas a instância mais provável.

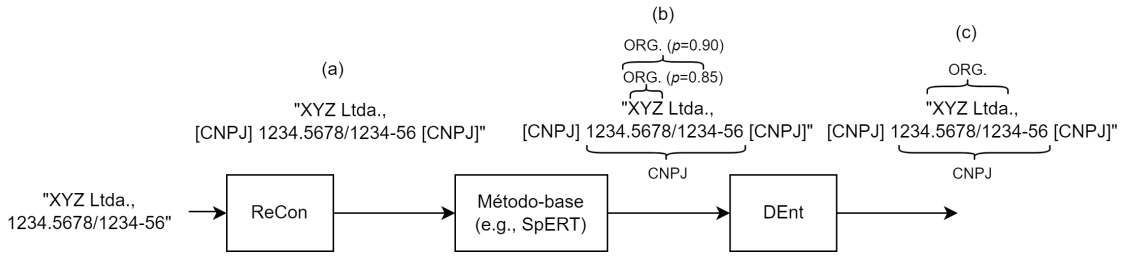

Figura 2. Etapas e exemplo da estratégia de processamento REN+ER proposta.

Tabela 1. Coleções de dados utilizadas

	LENER-BR	DO-MG	DO-MG-sintético
Tipo de texto	processos judiciais	diários oficiais	frases sintéticas
#frases com rótulos	10000	214	13000
#menções a entidades / #relações	9800 / -	1255 / 817	67000 / 50000
#tipos de entidade / #tipos de relação	6 / -	10 / 10	10 / 10

5. Metodologia de Avaliação

Nesta seção, são apresentadas as coleções de dados (Subseção 5.1), as métricas utilizadas na avaliação experimental (Subseção 5.2) e a parametrização do SpERT (Subseção 5.3).

5.1. Coleções de Dados

Para avaliar a eficácia das estratégias propostas, foram utilizadas duas coleções de dados reais: o conjunto de dados de processos judiciais LENER-BR [Luz de Araujo et al. 2018] e os Diários Oficiais do Estado de Minas Gerais (DO-MG) do ano de 2016, para os quais foi rotulada uma amostra representativa de 214 frases. Adicionalmente, foi gerada uma coleção de dados sintéticos com padrões similares ao DO-MG que foi utilizada para auxiliar no treinamento do SpERT (mas não para avaliá-lo). As principais estatísticas dessas coleções são sumarizadas na Tabela 1.

A amostra de 214 frases representativas e diversificadas do DO-MG foi selecionada através de uma estratégia de aprendizado ativo [Wang et al. 2021]. Com base nessa amostra, foram identificados 10 tipos relevantes de entidade e 10 tipos de relação presentes em diários oficiais. Os rótulos que definem os tipos de entidade são: PESSOA, ORGANIZAÇÃO, DATA, LOCAL, COMPETÊNCIA (cargos da administração pública), LEGISLAÇÃO, No_ATO, CPF, CNPJ e MASP (Matrícula de Servidor Público). Mais detalhes sobre esses rótulos, bem como sobre os tipos de relação entre esses tipos de entidade podem ser encontrados no repositório de dados disponibilizado.

Cada exemplo da amostra selecionada foi inspecionada por pelo menos três alunos dos programas de graduação e pós-graduação em Ciência da Computação da UFMG, os quais foram devidamente instruídos em relação a como a rotulação deveria ser feita. Prevaleceu, como gabarito para cada sequência de palavras, o rótulo mais votado. Houve discordância em menos de 10% das anotações tanto de entidades quanto de relações.

Para contornar a limitação da quantidade de dados rotulados na coleção DO-MG, foram gerados (apenas para fins de treino dos algoritmos e não para a sua avaliação) dados sintéticos (frases aleatórias que seguem determinados *templates* relacionando pares de entidades). Um exemplo de *template* para a relação local_residência, seria: "[PESSOA] *mora / reside / é residente em* [LOCAL]", onde [PESSOA] e [LOCAL] são substituídos por entidades aleatórias[2] correspondentes a esses tipos de entidade e uma das três expressões em itálico é selecionada aleatoriamente. Tais frases sintéticas, embora não constituam relações reais, apresentam padrões e estruturas úteis para o aprendizado do modelo de reconhecimento de entidades e relações.

Foram disponibilizados (vide *link* na Seção 1) tanto o código das estratégias quanto as coleções de dados sintéticos e reais, para garantir a reprodutibilidade dos re-

[2]Foram utilizadas entidades do tipo PESSOA e ORGANIZAÇÃO extraídas da coleção LENER-BR, bem como números de CPF, CNPJ, datas e números de atos aleatórios.

sultados e contribuir para o futuro desenvolvimento e avaliação de novas estratégias de REN e ER.

5.2. Métricas de Avaliação

Foram utilizadas as métricas Precisão, Revocação e F1, que capturam diferentes aspectos da eficácia do reconhecimento de entidades e de relações. Considerando x um tipo de entidade ou relação (e.g., x = PESSOA ou x = local_residência), a Precisão do algoritmo para reconhecer entidades ou relações do tipo x é calculada como:

$$\text{Precisão}(x) = \frac{\text{Número de acertos}}{\text{Total de vezes que o algoritmo reconheceu o tipo } x}$$

Já a Revocação mostra o quanto o algoritmo conseguiu cobrir as menções a entidades de um tipo x (ou relações do tipo x):

$$\text{Revocação}(x) = \frac{\text{Número de acertos}}{\text{Total de entidades ou relações do tipo } x \text{ mencionadas no texto}}$$

Por fim, o F1(x) é definido como a média harmônica entre Precisão(x) e Revocação(x), de modo a penalizar o valor dessa métrica quando qualquer uma dessas duas medidas (ou ambas) for baixa.

Para agregar as medidas de eficácia para todos os tipos de entidade e relação, são extraídas, para cada métrica, médias *micro* e *macro*. As medidas macro consistem na média das medidas sobre todos os tipos de entidade/relação. As medidas micro consideram a eficácia geral do algoritmo independentemente da eficácia por tipo de entidade/relação. Enquanto a média macro considera todos os tipos de entidade (ou de relação) igualmente importantes, sem negligenciar tipos mais infrequentes, a média micro tende a dar mais peso para tipos de entidade/relação mais frequentes.

Para fins de avaliação das estratégias propostas, as coleções de dados são divididas em três partições, denominadas conjuntos de *treino*, *validação* e *teste*. O treino é utilizado para o aprendizado do modelo de REN+ER, enquanto a validação é utilizada para ajuste de parâmetros. Por fim, os resultados são reportados no conjunto de teste. Para a coleção LENER-BR, foram utilizados os mesmos conjuntos de treino, validação e teste disponibilizados pelos autores. Para a coleção DO-MG, o conjunto de teste é formado por dados reais (metade das frases manualmente rotuladas), enquanto os conjuntos de validação e treino são formados por dados reais (a outra metade dos dados rotulados) e sintéticos. As vantagens dessa expansão do treino são discutidas na Subseção 6.1.

5.3. Parametrização do SpERT

Para parametrização da estratégia SpERT, foram utilizados os valores recomendados pelos seus autores [Eberts & Ulges 2020], a saber: taxa de aprendizado $l_r = 5 \times 10^{-5}$, número de épocas de treinamento $t = 20$, número de exemplos negativos por frase n^- = 100 (tanto para entidades como para relações) e tamanho de cada *batch* b_s = 2.

6. Resultados Experimentais

Nesta seção, são mostrados inicialmente os resultados da ampliação do conjunto de dados de treino da coleção DOMG utilizando frases sintéticas (Subseção 6.1) e, em seguida, os resultados das estratégias de pré-processamento (ReCon) e pós-processamento (DEnt) propostas (Subseção 6.2).

Tabela 2. Macro valores de Precisão, Revocação e F1 (e intervalos de 95% de confiança) com dados de treino reais apenas e com dados de treino expandidos com frases sintéticas. Melhores resultados (e empates estatísticos) em negrito.

	Entidades			Relações		
Dados de treino	Precisão	Revocação	F1	Precisão	Revocação	F1
Reais apenas	0.570 ± 0.027	$\mathbf{0.857 \pm 0.019}$	0.671 ± 0.026	0.325 ± 0.032	0.657 ± 0.033	0.392 ± 0.033
Reais+Sintéticos	$\mathbf{0.688 \pm 0.026}$	0.804 ± 0.022	$\mathbf{0.734 \pm 0.024}$	$\mathbf{0.588 \pm 0.034}$	$\mathbf{0.755 \pm 0.029}$	$\mathbf{0.657 \pm 0.033}$

6.1. Ampliação dos Dados de Treino com Frases Sintéticas

A Tabela 2 mostra os resultados das médias macro para as métricas precisão, revocação e F1 obtidas pela estratégia SpERT na coleção DOMG com e sem a ampliação por dados sintéticos. Os resultados para as médias micro são similares, sendo aqui omitidos por questões de espaço. Os melhores resultados, bem como os empates estatísticos de acordo com um teste-t bilateral com $\rho < 0.05$) são destacados em negrito.

Nota-se que o acréscimo de dados de treino sintéticos permitiu ganhos de 21% em precisão e 9% em F1 no reconhecimento de entidades (REN), bem como ganhos de 81% em precisão e 68% em F1 no reconhecimento de relações (ER). A única métrica em que não houve ganho foi a Macro-Revocação no caso do REN, o que é explicado pelo possível ruído introduzido pelos dados sintéticos.

Os ganhos bastante altos, particularmente para a tarefa ER, ocorrem porque o SpERT, por se tratar de um método supervisionado, depende de uma quantidade significativa de exemplos de treino rotulados para que tenha uma boa eficácia. Porém, as amostras de dados rotulados reais são reduzidas devido ao alto custo do processo manual de rotulação, em especial quando também é necessário identificar relações. Dessa forma, os resultados obtidos mostram os benefícios da expansão do treino com a estratégia de geração de frases sintéticas, que tem baixo custo e requer apenas algum conhecimento do domínio alvo ou a existência de *templates* de documentos desse domínio.

6.2. Eficácia das Estratégias Propostas

Nesta subseção, são apresentados os resultados de eficácia no reconhecimento de entidades e relações com e sem a aplicação das técnicas de pré-processamento (ReCon) e pós-processamento (DEnt) propostas. As Tabelas 3 e 4 mostram médias macro e micro, respectivamente, para as métricas precisão, revocação e F1 para a coleção de dados DOMG. A Tabela 5 mostra essas medidas para o reconhecimento de entidades na coleção LENER-BR. Para cada tabela, os melhores resultados (e empates estatísticos de acordo com um teste-t bilateral com $\rho < 0.05$) são destacados em negrito.

Nota-se que foram aplicadas ambas as estratégias ReCon e DEnt à coleção DOMG, enquanto apenas a DEnt foi aplicada à coleção LENER-BR. Isso foi feito porque essa última não apresenta rótulos para entidades regulares, como CPFs e telefones, o que dispensa o seu enriquecimento semântico através do ReCon. A seguir, são discutidos os resultados referentes à aplicação isolada de ReCon e DEnt (Subseções 6.2.1 e 6.2.2, respectivamente), bem como os resultados da aplicação conjunta de ambas as técnicas (Subseção 6.2.3).

Tabela 3. Macro-Precisão, Macro-Revocação e Macro-F1 (e intervalos de 95% de confiança) com e sem a aplicação das técnicas ReCon e DEnt, na coleção DO-MG. Melhores resultados (e empates estatísticos) em negrito.

		Entidades			Relações		
ReCon?	DEnt?	Precisão	Revocação	F1	Precisão	Revocação	F1
Não	Não	0.688 ± 0.026	0.804 ± 0.022	0.734 ± 0.024	0.588 ± 0.034	0.755 ± 0.029	0.657 ± 0.033
Não	Sim	0.796 ± 0.022	0.774 ± 0.023	0.781 ± 0.023	$\mathbf{0.733 \pm 0.030}$	0.714 ± 0.031	0.709 ± 0.031
Sim	Não	0.672 ± 0.026	$\mathbf{0.838 \pm 0.020}$	0.742 ± 0.024	0.518 ± 0.034	$\mathbf{0.786 \pm 0.028}$	0.616 ± 0.033
Sim	Sim	$\mathbf{0.820 \pm 0.021}$	$\mathbf{0.823 \pm 0.021}$	$\mathbf{0.820 \pm 0.021}$	$\mathbf{0.726 \pm 0.031}$	$\mathbf{0.770 \pm 0.029}$	$\mathbf{0.741 \pm 0.030}$

Tabela 4. Micro-Precisão, Micro-Revocação e Micro-F1 (e intervalos de 95% de confiança) com e sem a aplicação das técnicas de pré- e pós-processamento, na coleção DO-MG. Melhores resultados (e empates estatísticos) em negrito.

		Entidades			Relações		
ReCon?	DEnt?	Precisão	Revocação	F1	Precisão	Revocação	F1
Não	Não	0.705 ± 0.025	0.858 ± 0.019	0.774 ± 0.023	0.597 ± 0.034	0.788 ± 0.028	0.680 ± 0.032
Não	Sim	0.827 ± 0.021	0.831 ± 0.021	0.829 ± 0.021	$\mathbf{0.789 \pm 0.028}$	0.759 ± 0.029	0.774 ± 0.029
Sim	Não	0.720 ± 0.025	$\mathbf{0.892 \pm 0.017}$	0.797 ± 0.022	0.536 ± 0.034	$\mathbf{0.819 \pm 0.026}$	0.648 ± 0.033
Sim	Sim	$\mathbf{0.856 \pm 0.019}$	$\mathbf{0.872 \pm 0.019}$	$\mathbf{0.864 \pm 0.019}$	$\mathbf{0.797 \pm 0.028}$	$\mathbf{0.804 \pm 0.027}$	$\mathbf{0.801 \pm 0.027}$

6.2.1. Eficácia do ReCon

Comparando os resultados da primeira com a terceira linha das Tabelas 3 e 4, observa-se que o ReCon, quando aplicado isoladamente (sem a etapa de pós-processamento DEnt), gera ganhos modestos de 4% em revocação (para ambas médias macro e micro, tanto para o reconhecimento de entidades - REN - quanto para o de relações - ER), e não gera melhorias para a precisão. No entanto, quando aplicado juntamente com a DEnt, o ReCon gerou ganhos em todas as métricas de avaliação, como será discutido na Subseção 6.2.3.

Tais ganhos ocorrem porque o ReCon leva ao enriquecimento semântico das frases, o que auxilia o SpERT na identificação de tipos de entidade como CPFs e CNPJs, que são compostas basicamente apenas por números, havendo pouca semântica associada[3]. Nesses casos, o uso de expressões regulares tende a ser uma solução mais eficaz. Mesmo em casos de entidades mais irregulares (por exemplo, nomes de pessoas), houve melhorias significativas, pois muitas vezes elas ocorrem próximas a entidades regulares (por exemplo, um CPF muitas vezes ocorre próximo ao nome da pessoa física a ele associada, enquanto um CNPJ tende a ocorrer próximo ao nome da organização associada).

6.2.2. Eficácia da DEnt

Quando aplicada isoladamente, a DEnt contribui para um aumento de até 16% em macro-precisão para REN e um aumento de até 24% em macro-precisão para ER. Os ganhos correspondentes para micro-precisão são 17% e 32%. Quanto à revocação, como era esperado, não há ganhos, pois a DEnt apenas elimina menções menos prováveis e que ficaram sobrepostas a outras menções na saída. Como consequência dos ganhos relativamente altos em precisão, com perdas desprezíveis em revocação, há também ganhos em

[3]Muitas vezes, siglas como CPF e CNPJ já acompanham os números correspondentes a esses identificadores no texto. Apesar disso, em alguns casos o SpERT teve dificuldade de delimitar esses números, devido às variações de espaços e de pontuações entre eles. As expressões regulares ajudaram a extrair um número maior de casos de menções a esses tipos de entidade e o pós-processamento, como será discutido adiante, ajudou a delimitá-las de forma mais precisa.

Tabela 5. Micro e Macro valores de Precisão, Revocação e F1 (e intervalos de 95% de confiança) com e sem a aplicação da técnica de pós-processamento, na coleção LENER-BR. Melhores resultados (e empates estatísticos) em negrito.

	Médias-Micro			Médias-Macro		
DEnt?	Precisão	Revocação	F1	Precisão	Revocação	F1
Não	0.834 ± 0.007	$\mathbf{0.863 \pm 0.007}$	0.847 ± 0.007	0.848 ± 0.007	$\mathbf{0.866 \pm 0.007}$	0.857 ± 0.007
Sim	$\mathbf{0.848 \pm 0.007}$	0.855 ± 0.007	$\mathbf{0.850 \pm 0.007}$	$\mathbf{0.863 \pm 0.007}$	0.858 ± 0.007	$\mathbf{0.861 \pm 0.007}$

Macro-F1 (6% para REN e 8% para ER) e em Micro-F1 (7% para REN e 14% para ER).

Os maiores ganhos foram observados na coleção DO-MG. Na coleção LENER-BR, a DEnt promove ganhos modestos, mas estatisticamente significativos, de até 2% em macro e micro-precisão. Isso ocorre devido à menor complexidade da coleção LENER-BR, que tem apenas seis tipos de entidade e nenhuma relação rotulada. Para coleções mais complexas como o DO-MG, as estratégias propostas possuem um grande potencial de melhoria dos resultados.

6.2.3. Eficácia e Eficiência da Aplicação Conjunta das Estratégias ReCon e DEnt

Comparando o resultado da aplicação conjunta das estratégias ReCon e DEnt (quarta linha das Tabelas 3 e 4) com o resultado da estratégia SpERT original (primeira linha das mesmas tabelas), nota-se ganhos tanto em precisão (na faixa de 19% a 34% de ganho) como em revocação (pelo menos 2% de ganho). Esse resultado se deve ao aumento (modesto) de revocação oferecido pelo ReCon, e à melhoria (alta) da precisão oferecida pela DEnt, como discutido anteriormente. Portanto, o ReCon permite uma cobertura maior das menções às entidades (e consequentemente das relações), embora possa gerar resultados sobrepostos (imprecisão na delimitação das entidades, um problema que existe com ou sem pré-processamento). Esse problema, por sua vez, pode ser resolvido com a aplicação da técnica de pós-processamento DEnt.

Em relação à eficiência das estratégias propostas, nota-se que o custo adicional delas em relação à estratégia SpERT é desprezível. O tempo total de resposta do ReCon para todas as 214 frases de teste foi de 0,04s, o que equivale a 1,3% do tempo de execução do SpERT, que foi de 3,06s. O tempo de resposta da DEnt, também desprezível, equivale a menos de 1% do tempo de execução do SpERT. Tais resultados foram obtidos em um processador AMD Ryzen 5 5600X de 6 núcleos com 64GB de RAM.

7. Conclusões e Trabalhos Futuros

Neste artigo, foram propostas estratégias de Reforço Contextual (ReCon) e Delimitação de Entidades (DEnt) para a tarefa de Reconhecimento de Entidades e Relações (REN+ER). Tais estratégias são largamente aplicáveis a diversos tipos de dados não-estruturados. Dentre esses dados, destacam-se os documentos oficiais, que apresentam um conjunto rico de entidades que podem ser reconhecidas a partir de expressões regulares. A estratégia ReCon realiza uma marcação preliminar de entidades regulares no texto, permitindo a identificação de entidades que até mesmo um algoritmo do estado-da-arte como o SpERT pode não conseguir capturar sem este tipo de pré-processamento. A estratégia DEnt realiza um pós-processamento do resultado que visa unificar menções que ficaram sobrepostas na saída. Isso permitiu delimitar as entidades e relações de forma até 32% mais precisa em relação à estratégia do estado-da-arte. Ambas as

estratégias apresentam um custo adicional desprezível em relação ao custo total da tarefa. Como trabalhos futuros, pretende-se avaliar a eficácia das estratégias propostas quando aplicadas a outras estratégias de REN+ER, bem como a outras coleções de dados. Também pretende-se analisar em que situações é possível permitir um maior nível de sobreposição de entidades sem prejudicar a precisão.

Agradecimentos

Gostaríamos de agradecer o apoio do Ministério Público de Minas Gerais, por meio do projeto Capacidades Analíticas, bem como à CAPES, CNPq e FAPEMIG pelo apoio individual aos autores.

Referências

Brunner, U. & Stockinger, K. (2020). Entity Matching with Transformer Architectures - A Step Forward in Data Integration. In *International Conference on Extending Database Technology*, pages 463–473.

Caputo, A., Basile, P., & Semeraro, G. (2009). Boosting a Semantic Search Engine by Named Entities. In *Foundations of Intelligent Systems*, pages 241–250.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Eberts, M. & Ulges, A. (2020). Span-based Joint Entity and Relation Extraction with Transformer Pre-training. In *24th European Conference on Artificial Intelligence*, pages 2006–2013.

Eberts, M. & Ulges, A. (2021). An End-to-end Model for Entity-level Relation Extraction using Multi-instance Learning. In *Association for Computational Linguistics*, pages 3650–3660.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. In *Annual Meeting of the Association for Computational Linguistics*, pages 363–370.

Fu, J., Huang, X., & Liu, P. (2021). SpanNER: Named Entity Re-/Recognition as Span Prediction. In *Annual Meeting of the Association for Computational Linguistics*, pages 7183–7195.

Liu, C., Fan, H., & Liu, J. (2021). Span-based nested named entity recognition with pretrained language model. In Jensen, C. S., Lim, E.-P., Yang, D.-N., Lee, W.-C., Tseng, V. S., Kalogeraki, V., Huang, J.-W., & Shen, C.-Y., editors, *Database Systems for Advanced Applications*, pages 620–628.

Luz de Araujo, P. H., de Campos, T. E., de Oliveira, R. R. R., Stauffer, M., Couto, S., & Bermejo, P. (2018). LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In *International Conference on the Computational Processing of Portuguese (PROPOR)*, pages 313–323.

Niu, F., Zhang, C., Ré, C., & Shavlik, J. W. (2012). DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. *VLDS*, 12:25–28.

Patil, N., Patil, A., & Pawar, B. (2020). Named Entity Recognition using Conditional Random Fields. *Procedia Computer Science*, 167:1181–1188.

Silva, L., Canalle, G. K., Salgado, A. C., Lóscio, B., & Moro, M. (2019). Uma Análise Experimental do Impacto da Seleção de Atributos em Processos de Resolução de Entidades. In *SBBD*, pages 37–48.

Wang, T., Zhao, X., Lv, Q., Hu, B., & Sun, D. (2021). Density weighted diversity based query strategy for active learning. In *IEEE International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 156–161.

Zhang, S., He, L., Vucetic, S., & Dragut, E. (2018). Regular Expression Guided Entity Mention Mining from Noisy Web Data. In *Empirical Methods in Natural Language Processing*, pages 1991–2000.