An Annotated Dataset for Automatic Extraction of Entities and Restrictions from Business Process Models

Diogo S. Candido[1], João Victor Berti Lima[2], Hilário Oliveira[1], Mateus B. Costa[1]

[1]Programa de Pós-graduação em Computação Aplicada (PPComp)
Instituto Federal do Espírito Santo (IFES) – Campus Serra

[2]Coordenadoria do Curso de Engenharia de Controle e Automação
Instituto Federal do Espírito Santo (IFES) – Campus Serra
Av. dos Sabiás, 330 – Morada de Laranjeiras – 29.166-630 – Serra – ES – Brasil

diogosantanaime@gmail.com, joaovictorbertilima@gmail.com,
{hilario.oliveira,mcosta}@ifes.edu.br

Abstract. *Business Process Modeling is often perceived as a high-potential activity that is difficult to implement. Various techniques and methods have been proposed and investigated to support this activity, with emphasis on the use of natural language processing techniques. However, the scarcity of datasets specifically for this purpose constitutes an important limitation recognized by the literature. This work proposes an annotated dataset for identifying typical business process entities and restrictions. Experiments conducted focusing on entity recognition suggest that the BiLSTM-CRF architecture, with word embeddings extracted from the GloVe, Flair, and BERT models, achieved the best performance based on the micro average of the f1-score measure.*

Resumo. *A Modelagem de Processos de Negócio é frequentemente percebida como uma atividade com alto retorno potencial, mas de difícil realização. Diversas técnicas e métodos têm sido propostos e investigados para apoiar esta atividade, destacando-se o uso de técnicas de processamento de linguagem natural. Entretanto, a escassez de conjuntos de dados especificamente para este fim constitui uma importante limitação reconhecida pela literatura. Este trabalho propõe uma base de dados anotada para a identificação de entidades e restrições típicas de processos de negócios. Experimentos conduzidos com foco no reconhecimento de entidades sugerem que a arquitetura BiLSTM-CRF, com incorporações de palavras extraídas dos modelos GloVe, Flair e BERT, alcançou o melhor desempenho com base na média micro da medida f1-score.*

1. Introdução

A criação manual de um modelo de processo a partir de dados textuais pode ser uma tarefa demorada, que demanda muitos recursos [Bellan et al. 2020]. Este cenário desafiador normalmente ocorre porque o conhecimento necessário para interpretar os conteúdos relevantes reside nos usuários que executam ativamente esses processos, os quais muitas vezes não possuem as habilidades necessárias para formalizar modelos de processos de negócio [Van der Aa et al. 2019]. Além disso, apesar

dos avanços nas técnicas de automação de processos e extração de modelos, como a mineração de processos, a modelagem de processos corporativos continua sendo predominantemente uma tarefa manual que requer um esforço humano significativo [Shilov et al. 2023]. Consequentemente, a adoção de métodos automatizados em tarefas de modelagem tem sido identificada como um desafio na gestão de processos de negócio [Beerepoot et al. 2023].

A abundância de dados textuais oferece aos modeladores a oportunidade de acesso a uma grande fonte de informações, que pode ser usada na transcrição e elaboração destes modelos. Este contexto, por sua vez, apresenta um novo desafio, pois os modeladores frequentemente se deparam com dificuldades, tanto cognitivas quanto operacionais, para analisar manualmente este tipo de fonte [Shilov et al. 2023]. Neste contexto, técnicas de Processamento de Linguagem Natural (PLN) têm sido investigadas visando mitigar os obstáculos cognitivos e a sobrecarga gerada pelo grande volume de trabalho. Resultados recentes no contexto da modelagem de processos sugerem dificuldades, de técnicas de PLN empregadas, de lidar com diversos aspectos de descrições desses modelos, tais como, níveis linguísticos, ambiguidades, limitações de conjuntos de dados e a alta complexidade do controle de fluxo [Bellan et al. 2022b]. Tais dificuldades prejudicam sua eficácia, especialmente quando lidam com descrições de processos que apresentam estruturas de controle de fluxo complexas como, *AND-split*, *AND-join*, *OR-split*, *OR-join* e ciclos, comumente encontradas em processos de negócios do mundo real [Deng et al. 2016].

Neste artigo, apresentamos um conjunto de dados anotados por especialistas humanos, composto por 133 descrições de processos de negócios escritas em inglês, coletadas de diversas fontes de informação [Friedrich et al. 2011, Mangler and Klievtsova 2023, Dumas et al. 2018]. A base de dados possui anotações de entidades de interesse, como atores e atividades, bem como relações de restrição entre essas entidades. Dois avaliadores humanos, seguindo uma abordagem de modelagem declarativa, que especifica restrições e regras sobre o comportamento do processo sem definir uma sequência rígida de atividades, avaliaram mais de 1.300 sentenças. O objetivo do conjunto de dados é ser um ferramental para treinar algoritmos voltados para uma análise preliminar desses textos. Esta análise visa extrair as entidades e as relações de restrição lógico temporais e de comprometimento de atores para com atividades e eventos. Os elementos (entidades e restrições) extraídos podem ser utilizados na construção automatizada de modelos processos utilizando, por exemplo, notações procedurais como, a notação de modelagem de processos de negócio, do inglês *Business Process Modeling Notation* (BPMN) [Barba et al. 2013].

Experimentos foram conduzidos para demonstrar a viabilidade do conjunto de dados apresentado e estabelecer bases para comparações futuras. Inicialmente, foi priorizada a tarefa de Reconhecimento de Entidades Nomeadas (REN) utilizando incorporações de palavras e algoritmos de aprendizado profundo para identificar e classificar os elementos dos processos descritos nos textos. Os resultados dos experimentos demonstraram que a abordagem utilizando uma arquitetura de rede neural do tipo *Bidirectional Long Short-Term Memory Network* (BiLSTM) em conjunto com o algoritmo *Conditional Random Fields* (CRF), aqui chamada de BiLSTM-CRF, combinada com representações de incorporação de palavras extraídas dos mo-

delos GloVe, Flair e BERT, obteve o melhor desempenho com base na média micro da medida *f1-score*.

O conjunto de dados e o código-fonte utilizado nos experimentos estão disponíveis: `https://github.com/laicsiifes/bpm_dataset`.

2. Fundamentos

2.1. Modelagem do Controle de Fluxo de Processos de Negócio

A perspectiva de controle de fluxo refere-se à especificação formal de dependências que governam a sequência de atividades em um processo [Fionda and Guzzo 2020]. Portanto, sua modelagem fornece a especificação de regras que determinam o comportamento do processo ou sua semântica operacional. Linguagens e notações para tal propósito podem ser categorizadas como declarativas (baseadas em restrições) e procedurais (imperativas). Enquanto as notações declarativas visam fornecer restrições limitantes para a execução do processo, as notações procedurais visam estabelecer os caminhos específicos (agendas de execução) que o processo deve seguir. As notações procedurais possuem elementos de modelo baseados na Álgebra de Processos, cujos operadores permitem a especificação de ordem (.), escolha (+) e paralelismo (|). A Figura 1 ilustra o uso desses operadores para descrever um modelo simples de processamento de pedidos. Nesta descrição, o operador "." estabelece ordem temporal entre as tarefas. Por exemplo, "Novo Pedido" deve ocorrer antes de "Registrar Pedido". O operador "+" estabelece uma relação de escolha entre os conjuntos de operandos, indicando que apenas um dos conjuntos deve ser executado em uma dada instância de execução. O Operador "|" estabelece que os conjuntos de operandos devem ser executados em qualquer ordem ou simultaneamente.

$$Proc = \text{“Novo Pedido”.“Registrar Pedido”.“Verificar Estoque”.} \begin{pmatrix} \text{(“Produto Disponível”.(“Enviar Produto”|“Cobrar”))} \\ + \text{(“Produto Indisponível”.“Cancelar Pedido”)} \end{pmatrix}$$

Figura 1. Modelo de Processamento de Pedido usando Álgebra de Processos.

Linguagens e notações de processos declarativos são inspiradas em linguagens baseadas em lógica, como a Lógica Temporal Linear (LTL) [Fionda and Guzzo 2020]. A LTL é baseada na formulação de sentenças que restringem lógica e temporalmente o comportamento das variáveis consideradas. Dessa forma, é possível impor a condição de existência (execução) de atividades com base na existência de outras atividades. A linguagem incorpora operadores booleanos e operadores modais temporais, como X para "*next*" (próximo), F para "*finally*" (finalmente) e R para "*release*". Um processo pode ser modelado com LTL criando um conjunto de sentenças LTL. A Figura 2 ilustra o processamento de pedidos em LTL.

Sentenças escritas na LTL podem ser desordenadas e não relacionadas entre si. Dessa forma, esses modelos têm a vantagem de poderem ser construídos de maneira fragmentada, por exemplo, em momentos diferentes ou em espaços organizacionais distintos. Sendo uma descrição declarativa restritiva e não específica e modelos procedurais sendo específicos e rigidamente definidos, modelos procedurais

$$\text{Novo Pedido} \rightarrow \mathbf{X}(\text{Registrar Pedido})$$
$$\text{Registrar Pedido} \rightarrow \mathbf{X}(\text{Verificar Estoque})$$
$$\text{Verificar Estoque} \rightarrow \mathbf{X}(\text{Produto Disponível} \vee \text{Produto Indisponível})$$
$$\text{Produto Disponível} \rightarrow \mathbf{X}(\text{Enviar Produto} \vee \text{Cobrar})$$
$$\text{Produto Indisponível} \rightarrow \mathbf{X}(\text{Cancelar Pedido})$$
$$\neg(\text{Enviar Produto} \wedge \text{Cancelar Pedido})$$
$$\neg(\text{Cobrar} \wedge \text{Cancelar Pedido})$$

Figura 2. Modelo de Processo de Pedido em LTL.

podem ser obtidos a partir de descrições declarativas. No exemplo utilizado, a descrição do Processamento de Pedidos em Álgebra de Processos é dita consistente com a sua especificação em LTL.

2.2. Notação de Modelagem Baseada em Situações

Na análise de dados textuais, visando construir processos, identificar todos os elementos envolvidos no modelo, bem como estabelecer uma sentença completa que descreva o controle de fluxo do mesmo, é uma tarefa complexa. Nem sempre o processo está claramente descrito, livre de ambiguidades e incompletudes. Dessa forma, modelos procedurais são difíceis de se obter. Por outro lado, a obtenção de modelos declarativos pode ser mais bem sucedida. Tais modelos podem, posteriormente, ser usados para derivar modelos procedurais específicos.

Dessa forma, este trabalho considerou a obtenção indireta de modelos, buscando em um primeiro passo a identificação de entidades e de restrições entre essas entidades. Para tanto, foi utilizada como base para o conjunto de tipos de restrições a serem identificadas, uma Notação de Modelagem declarativa Baseada em Situações (NMBS) [Costa and Tamzalit 2017]. Foram consideradas as seguintes restrições da NMBS:

Dependência. Um conjunto de Objetos de Fluxo Ativo (AFOs, do inglês *Active Flow Objects*) com uma dependência de execução temporal entre eles. As dependências são subclassificadas em dois tipos: *Estrita* ($\lhd$, DEP) e
Circunstancial ($\trianglelefteq$, DEPC). Em uma relação de dependência estrita, se b depende de a, então b só pode ser executado em um fluxo se e somente se a tiver sido executado antes. Em uma relação de dependência circunstancial, se b depende de a, então b pode ser executado em um fluxo onde a foi executado antes ou, em um fluxo onde a não é executado em nenhum momento.

Não-Coexistência. Um conjunto de objetos de fluxo com uma relação de não-coexistência ($\otimes$, XOR) no mesmo fluxo de execução, normalmente mapeada para uma relação XOR em notações procedurais.

União. Um conjunto de objetos de fluxo com uma relação de união ($\oplus$, UNI) no mesmo fluxo de execução.

Além dos elementos mencionados anteriormente, foi adicionado um operador que estabelece a relação de responsabilidade de execução ($perform$). Esses elementos e as entidades anotadas são discutidos em mais detalhes na Seção 4.2.

3. Trabalhos Relacionados

Bases de dados (*corpora*) anotados servem como referências para o treinamento e a validação de modelos, especialmente aqueles baseados em algoritmos de Aprendizado de Máquina (AM). Diversas iniciativas tratam dos modelos de processos de negócio por meio da extração de informações utilizando técnicas de AM e PLN, como os trabalhos de [Qian et al. 2020, Ackermann et al. 2021, López et al. 2021], sendo a base de dados PET [Bellan et al. 2022b] a que mais se aproxima da proposta deste trabalho. O conjunto de dados PET tem como objetivo servir de base para a obtenção automática de modelos procedurais na notação BPMN a partir de documentos textuais. Para isso, o PET inclui anotações para a rotulagem de atividades, desvios paralelos, desvios exclusivos, atores e sequências.

Em outras iniciativas, diversas abordagens para a tarefa de Reconhecimento de Entidades Nomeadas (REN), como *Conditional Random Fields* (CRF) e *Bidirectional Long Short-Term Memory Network* (BiLSTM), têm sido utilizadas para identificar e extrair elementos de processos de negócio a partir de textos [Bellan et al. 2023]. O algoritmo CRF pode ser usado para modelar sequências de palavras e melhorar a precisão da extração, enquanto o BiLSTM classifica entidades de interesse, como atividades, atores e condições [Li et al. 2020]. Quando combinada com o CRF em uma arquitetura BiLSTM-CRF, essa abordagem aproveita os pontos fortes de ambos os modelos, permitindo a extração de entidades complexas em descrições de processos de negócio [Li et al. 2020]. Algoritmos de incorporação de palavras (*word embeddings*), como GloVe [Pennington et al. 2014], Flair [Akbik et al. 2018] e BERT [Devlin et al. 2018], também são utilizados para capturar significados semânticos contextuais, proporcionando uma maior compreensão dos textos.

Estudos recentes incluem a extração de modelos declarativos com pipelines de PLN personalizados [Van der Aa et al. 2019] e a combinação de aprendizado profundo com regras para grafos de resposta a condições dinâmicas [López et al. 2021]. Outras técnicas, como grafos neurais [Ackermann et al. 2021], são usadas para análise semântica. Trabalhos adicionais concentram-se na extração de elementos específicos, utilizando abordagens baseadas em regras e técnicas de PLN [Ferreira et al. 2017, Epure et al. 2015], e métodos como classificação em múltiplos níveis [Qian et al. 2020] e uma linguagem de consulta para padrões baseados em árvore [Quishpi et al. 2020]. Pesquisas envolvendo o uso de grandes modelos de linguagem (LLMs, do inglês *Large Language Models*), como o GPT-3 [Bellan et al. 2022a], destacam o potencial e os desafios dessas tecnologias. A diversidade de técnicas e conjuntos de dados ressalta a necessidade de métricas padronizadas e bases de dados mais amplas.

4. Construção do Conjunto de Dados Anotados

4.1. Aquisição de Dados

Existe uma escassez de conjuntos de dados publicamente anotados disponíveis para modelagem de processos de negócio a partir de descrições textuais [Bellan et al. 2023]. Neste contexto, a construção do conjunto de dados começou com a coleta de textos, sem anotações prévias, que contêm descrições de processos de negócios em inglês.

Para isso, foram coletados os textos propostos por [Friedrich et al. 2011], que são amplamente utilizados no campo da pesquisa sobre extração de informações de processos de negócio descritos em linguagem natural. Outra fonte de textos utilizada é a proposta por [Klievtsova et al. 2023], que fornece 24 diferentes descrições textuais de processos de negócios. Para enriquecer ainda mais a coleção de textos, foram adicionados 48 textos de exercícios e exemplos de descrições de processos de negócios, obtidos de [Dumas et al. 2018]. Finalmente, foram adicionadas 16 descrições textuais adicionais de processos de negócios coletados pelo autor, compreendendo textos usados na disciplina de Modelagem de Processos de Negócio oferecida por um dos autores no nível de graduação. Vale ressaltar que os textos coletados não possuem nenhuma anotação prévia, sendo todos eles anotados neste trabalho.

A Tabela 1 lista as fontes de informação das quais as 133 descrições textuais usadas para construir o conjunto de dados proposto foram obtidas.

Tabela 1. Autores, origem e quantidade de textos no conjunto de dados do corpus.

Autores	**Origem**	**Textos**
[Friedrich et al. 2011]	Conjunto de dados original proposto por Friedrich et al.	45
[Mangler and Klievtsova 2023]	Novas descrições textuais de processos de negócios	24
[Dumas et al. 2018]	Exercícios e exemplos de descrições de processos de negócios	48
Coletado a partir de exercícios de aula	Disciplina de graduação em modelagem de processos de negócios	16
Total		**133**

4.2. Processo de Anotação

A anotação do conjunto de dados foi baseada na Notação de Modelagem Baseada em Situações discutida na Seção 2, a partir da qual as relações lógico-temporais de Dependência Estrita, Dependência Circunstancial, União e Não-Coexistência foram derivadas. Adicionalmente, foram incluídas relações de responsabilidade pelas ações do processo. Para esse propósito, a entidade denominada *Actor* (Ator) foi criada. As entidades que representam objetos de fluxo ativo foram categorizadas como *Activity* (Atividade), *Trigger* (Gatilho) e *Catch* (Captura). A entidade *Conditional* (Condicional), que também foi estabelecida, permite definir a condição associada a uma dependência circunstancial e a não-coexistência. Na Tabela 2 é apresentada uma breve descrição dos tipos de entidades e relações consideradas em nosso processo de anotação.

A anotação foi realizada por dois avaliadores humanos com experiência em modelagem de processos de negócios. A ferramenta Doccano[1], que permite a anotação de elementos textuais e é comumente empregada na criação de bases de dados no campo de PLN, foi adotada para auxiliar no processo de anotação. Um guia de anotação[2] foi desenvolvido para realizar a anotação de forma sistemática, permitir a

[1] `https://github.com/doccano/doccano`

[2] `https://github.com/laicsiifes/bpm_dataset`

Tabela 2. Tipos de entidades e relações consideradas no processo de anotação.

Entidades	
Nome	**Descrição**
Actor	Identifica a responsabilidade pelas ações dentro de um processo.
Activity	Identifica um objeto de fluxo ativo que realiza uma tarefa dentro de um processo.
Trigger	Identifica um evento que aciona algo dentro de um processo.
Catch	Identifica um evento que captura algo dentro do processo.
Conditional	Identifica a condição associada a uma dependência circunstancial.
Relações	
Strict Dependence	Identifica dependências estritas.
Circumstantial Dependence	Identifica dependências circunstanciais.
Union	Identifica relações de União.
Non-Co-Existence	Identifica relações de Não-coexistência.
Perform	Identifica a responsabilidade de um ator na realização de uma atividade.

participação de vários anotadores e garantir a consistência das anotações. Esse guia também visa fornecer os padrões e critérios adotados para a anotação dos textos, possibilitando a repetibilidade do trabalho e do conjunto de dados formado com os textos anotados.

Para ilustrar a anotação e seu resultado, considere a seguinte descrição de um processo de pedido de compra: "When a new product is requested, the supplier checks the inventory. If the product is available, it is labeled. After that, the product is sent to the dock. Otherwise, the product is requested from the manufacturer, and a delayed delivery notice is issued to the requester." As entidades e relações anotadas neste exemplo de descrição textual são apresentadas na Figura 3.

Após o processo de anotação realizado pelos dois anotadores, o mais experiente deles foi responsável por revisar cada uma das anotações. Esse processo foi baseado nos critérios propostos por [Yuan et al. 2021], que foram complementados com os critérios de ausência e ambiguidade, que identificam respectivamente a ausência de anotação e a presença de ambiguidade na anotação.

No final do processo de anotação, 1.361 sentenças foram anotadas pelos dois anotadores. A Tabela 3 apresenta estatísticas descritivas do número total de entidades e relações identificadas. O processo de anotação resultou em 4.625 anotações de entidades e relações. Entre elas, as categorias *Actor* e *Activity* têm as maiores frequências, com 1.299 e 1.252 anotações, respectivamente, indicando um forte foco na identificação dos principais participantes e suas ações dentro dos textos. Dependência Estrita e Condicional também apresentam um número significativo de menções, refletindo a ênfase na captura das dependências e condições nos processos.

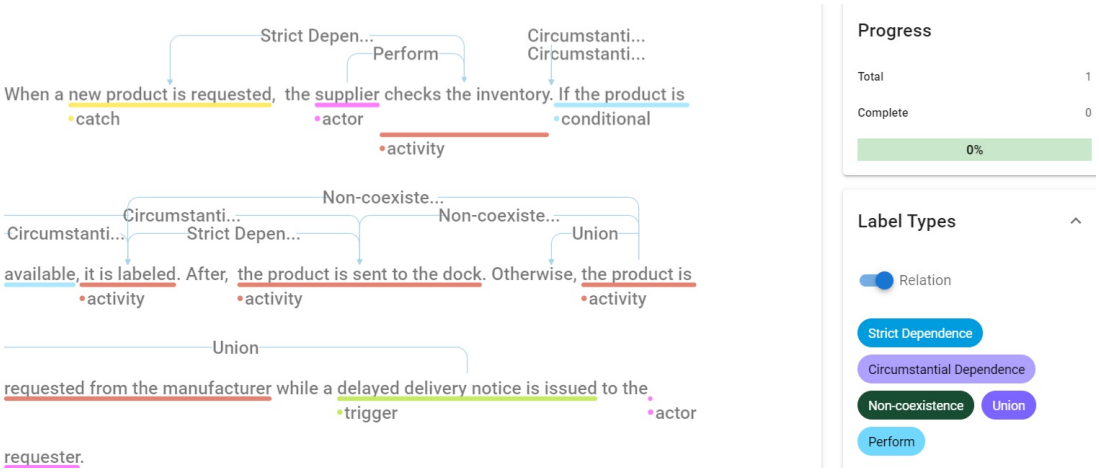

Figura 3. Exemplo de anotação para um Processo de Pedido de Compra.

Tabela 3. Número total de entidades e relações anotadas no conjunto de dados.

Nome	Total de Anotações
Entidades	
Actor	1.299
Activity	1.252
Trigger	158
Catch	181
Conditional	496
Relações	
Strict Dependence	642
Circumstantial Dependence	494
Union	58
Non-Co-Existence	45
Total	4.625

5. Experimentos

5.1. Configuração Experimental

Experimentos foram realizados para analisar a viabilidade do treinamento de modelos de aprendizado de máquina para identificar automaticamente as entidades nomeadas anotadas. O objetivo é fornecer resultados preliminares para este conjunto de dados, que podem servir de base para pesquisas futuras. É importante ressaltar que, embora a base de dados tenha anotações das relações entre as entidades, neste trabalho focamos apenas nas entidades nomeadas e deixamos a exploração das relações para análises futuras.

Os 133 documentos anotados foram convertidos para o formato *Conference on Natural Language Learning* (CoNLL), um padrão comumente utilizado para bancos de dados em tarefas de REN. O conjunto de dados foi então dividido em 107 documentos para treinamento, 13 para validação e 13 para teste. Duas abordagens para a tarefa de REN foram avaliadas.

A primeira abordagem utiliza características linguísticas das palavras, como

a classe gramatical das palavras, se começa letra maiúscula ou minúscula, entre outras, e treina o algoritmo CRF [do Amaral and Vieira 2014]. A segunda abordagem emprega uma rede neural do tipo BiLSTM usando uma última camada composta pelo algoritmo CRF, formando uma arquitetura comumente chamada de BiLSTM-CRF. Um componente essencial das arquiteturas de redes neurais para as tarefas de PLN é o modelo de incorporação de palavras (*word embeddings*) usado para representar as palavras. Para isso, avaliamos o modelo tradicional GloVe [Pennington et al. 2014] e incorporações contextuais extraídas dos modelos Flair [Akbik et al. 2018] e *Bidirectional Encoder Representations from Transformers* (BERT) [Devlin et al. 2018]. Inspirados por resultados promissores em trabalhos anteriores, avaliamos esses três modelos de representação tanto individualmente quanto combinados [da Silva and de Oliveira 2022].

Para o treinamento do modelo BiLSTM-CRF, usamos o framework Flair [Akbik et al. 2018], que oferece ferramentas para o treinamento de modelos para tarefas de classificação sequencial. O modelo foi treinado por no máximo 100 épocas com um tamanho de lote de 32, utilizando otimização por Descida de Gradiente Estocástica (SGD, do inglês *Stochastic Gradient Descent*) com uma taxa de aprendizado de 0,1 e uma paciência de 20 épocas.

5.2. Resultados e Discussão

Na Tabela 4 são apresentados os resultados experimentais baseados na medida de avaliação *f1-score* computada em nível de entidade, ou seja, considera-se apenas um resultado correto se o modelo estimar corretamente todas as palavras que compõem a entidade. Os melhores resultados para cada entidade estão destacados em negrito. Ao analisar os resultados para as entidades individualmente, observa-se que todos os modelos avaliados alcançaram alto desempenho (superior a $0,8$) no reconhecimento das entidades *Actor*. Observa-se também que um desempenho razoável foi alcançado para as entidades *Condition* e *Activity*, mas os resultados obtidos para as entidades *Trigger* e especialmente *Catch* foram muito baixos. A estratégia de combinar os três modelos de incorporação de palavras (GloVe + Flair + BERT) apresentou o melhor desempenho para a maioria dos rótulos e para a média geral micro, destacando como o uso representações combinadas de *word embeddings* pode melhorar a capacidade de capturar relações e dependências complexas dentro das descrições de processos.

Tabela 4. Resultados dos experimentos de REN usando a medida f1-score.

Entidades	**CRF**	**BiLSTM + CRF**					
		Glove (G)	**Flair (F)**	**BERT (B)**	**G + F**	**G + B**	**G + F + B**
Activity	0,463	0,400	0,552	0,579	0,588	0,496	**0,615**
Actor	0,804	0,839	0,912	0,902	0,915	0,900	**0,918**
Catch	0,095	0,077	0,000	0,167	0,100	**0,260**	0,220
Condition	0,694	0,645	0,727	0,765	0,727	0,712	**0,789**
Trigger	0,207	0,200	0,370	0,296	0,167	0,182	**0,414**
Média Micro	0,611	0,610	0,707	0,711	0,712	0,669	**0,744**

O baixo desempenho na identificação das entidades *Catch* e *Trigger* pode estar relacionado ao pequeno número de exemplos no conjunto de dados. Portanto,

os modelos não foram capazes de aprender boas regras de extração para essas entidades. Outra característica desafiadora que notamos em todas as categorias de entidades, exceto *Actor*, e especialmente em *Activity* e *Condition*, é que elas são compostas por várias palavras, geralmente três ou quatro. Tal cenário é desafiador para os modelos de REN, que precisam estimar corretamente os rótulos das entidades e quais palavras formam uma única menção de uma entidade (fronteiras das entidades).

6. Conclusão

Este trabalho apresentou a criação de uma base de dados de referência contendo 133 documentos que incluem entidades e restrições (relações) de interesse no contexto de modelagem de processo de negócio. Esses documentos foram anotados por especialistas humanos, resultando em um conjunto de dados composto por 1.361 sentenças e 4.625 elementos (entidades e relações) anotados. Experimentos iniciais, considerando a tarefa de reconhecimento de entidades nomeadas, avaliaram a base de dados proposta e definiram marcos iniciais. Os resultados experimentais demonstraram que a combinação das representações de incorporação de palavras extraídas do GloVe, Flair e BERT melhoraram a precisão do reconhecimento de entidades usando uma arquitetura do tipo BiLSTM-CRF.

Apesar dos resultados encorajadores obtidos, o estudo necessita de mais exploração para aprimorar as metodologias, com foco na expansão do conjunto de dados e no refinamento das anotações para evitar interferências na precisão dos modelos. Além disso, experimentos envolvendo as relações anotadas entre as entidades são essenciais para desenvolver um modelo de mapeamento de processos declarativos.

Agradecimentos

Os autores agradecem ao Ifes, apoio da FAPES e CAPES (processo 2021-2S6CD, nº FAPES 132/2021) por meio do PDPG (Programa de Desenvolvimento da Pós-Graduação, Parcerias Estratégicas nos Estados).

Referências

Ackermann, L., Neuberger, J., and Jablonski, S. (2021). Data-driven annotation of textual process descriptions based on formal meaning representations. In *International Conference on Advanced Information Systems Engineering*, pages 75–90. Springer.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pages 1638–1649.

Barba, I., Del Valle, C., Weber, B., and Jimenez, A. (2013). Automatic generation of optimized business process models from constraint-based specifications. *International Journal of Cooperative Information Systems*, 22(02):1350009.

Beerepoot, I., Di Ciccio, C., Reijers, H. A., Rinderle-Ma, S., Bandara, W., Burattin, A., Calvanese, D., Chen, T., Cohen, I., Depaire, B., et al. (2023). The biggest business process management problems to solve before we die. *Computers in Industry*, 146:103837.

Bellan, P., Dragoni, M., and Ghidini, C. (2020). A qualitative analysis of the state of the art in process extraction from text. *DP@AI*IA*, pages 19–30.

Bellan, P., Dragoni, M., and Ghidini, C. (2022a). Extracting business process entities and relations from text using pre-trained language models and in-context learning. In *International Conference on Enterprise Design, Operations, and Computing*, pages 182–199. Springer.

Bellan, P., van der Aa, H., Dragoni, M., Ghidini, C., and Ponzetto, S. P. (2022b). Pet: an annotated dataset for process extraction from natural language text tasks. In *International Conference on Business Process Management*, pages 315–321. Springer.

Bellan, P., van der Aa, H., Dragoni, M., Ghidini, C., and Ponzetto, S. P. (2023). Process extraction from text: Benchmarking the state of the art and paving the way for future challenges. *arXiv preprint arXiv:2110.03754*.

Costa, M. B. and Tamzalit, D. (2017). Recommendation patterns for business process imperative modeling. In *Proceedings of the Symposium on Applied Computing*, pages 735–742.

da Silva, M. G. and de Oliveira, H. T. A. (2022). Combining word embeddings for portuguese named entity recognition. In *International Conference on Computational Processing of the Portuguese Language*, pages 198–208. Springer.

Deng, S., Wang, D., Li, Y., Cao, B., Yin, J., Wu, Z., and Zhou, M. (2016). A recommendation system to facilitate business process modeling. *IEEE transactions on cybernetics*, 47(6):1380–1394.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

do Amaral, D. O. F. and Vieira, R. (2014). Nerp-crf: uma ferramenta para o reconhecimento de entidades nomeadas por meio de conditional random fields. *Linguamática*, 6(1):41–49.

Dumas, M., La Rosa, M., Mendling, J., Reijers, H. A., et al. (2018). *Fundamentals of business process management*, volume 2. Springer.

Epure, E. V., Martín-Rodilla, P., Hug, C., Deneckère, R., and Salinesi, C. (2015). Automatic process model discovery from textual methodologies. In *2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS)*, pages 19–30. IEEE.

Ferreira, R. C. B., Thom, L. H., and Fantinato, M. (2017). A semi-automatic approach to identify business process elements in natural language texts. In *International Conference on Enterprise Information Systems*, volume 2, pages 250–261. SCITEPRESS.

Fionda, V. and Guzzo, A. (2020). Control-flow modeling with declare: Behavioral properties, computational complexity, and tools. *IEEE Transactions on Knowledge & Data Engineering*, 32(05):898–911.

Friedrich, F., Mendling, J., and Puhlmann, F. (2011). Process model generation from natural language text. In *Advanced Information Systems Engineering: 23rd International Conference, CAiSE 2011, London, UK, June 20-24, 2011. Proceedings 23*, pages 482–496. Springer.

Klievtsova, N., Benzin, J.-V., Kampik, T., Mangler, J., and Rinderle-Ma, S. (2023). Conversational process modelling: State of the art, applications, and implications in practice. *arXiv preprint arXiv:2304.11065*.

Li, J., Sun, A., Han, J., and Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, 34(1):50–70.

López, H. A., Strømsted, R., Niyodusenga, J.-M., and Marquard, M. (2021). Declarative process discovery: Linking process and textual views. In Nurcan, S. and Korthaus, A., editors, *Intelligent Information Systems*, pages 109–117, Cham. Springer International Publishing.

Mangler, J. and Klievtsova, N. (2023). Textual process descriptions and corresponding bpmn models. https://doi.org/10.5281/zenodo.7783492.

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

Qian, C., Wen, L., Kumar, A., Lin, L., Lin, L., Zong, Z., Li, S., and Wang, J. (2020). An approach for process model extraction by multi-grained text classification. In *Advanced Information Systems Engineering: 32nd International Conference, CAiSE 2020, Grenoble, France, June 8–12, 2020, Proceedings 32*, pages 268–282. Springer.

Quishpi, L., Carmona, J., and Padró, L. (2020). Extracting annotations from textual descriptions of processes. In *Business Process Management: 18th International Conference, BPM 2020, Seville, Spain, September 13–18, 2020, Proceedings 18*, pages 184–201. Springer.

Shilov, N., Othman, W., Fellmann, M., and Sandkuhl, K. (2023). Machine learning for enterprise modeling assistance: an investigation of the potential and proof of concept. *Software and Systems Modeling*, 22(2):619–646.

Van der Aa, H., Di Ciccio, C., Leopold, H., and Reijers, H. A. (2019). Extracting declarative process models from natural language. In *Advanced Information Systems Engineering: 31st International Conference, CAiSE 2019, Rome, Italy, June 3–7, 2019, Proceedings 31*, pages 365–382. Springer.

Yuan, A., Ippolito, D., Nikolaev, V., Callison-Burch, C., Coenen, A., and Gehrmann, S. (2021). Synthbio: A case study in faster curation of text datasets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.