




















# Contextual Reinforcement, Entity Delimitation and Generative Data Augmentation for Entity Recognition and Relation Extraction in Official Documents

Fabiano Muniz Belém   [ Universidade Federal de Minas Gerais | [fmuniz@dcc.ufmg.br](mailto:fmuniz@dcc.ufmg.br) ]  
Claudio Valiense   [ Universidade Federal de Minas Gerais | [claudio.valiense@dcc.ufmg.br](mailto:claudio.valiense@dcc.ufmg.br) ]  
Celso França   [ Universidade Federal de Minas Gerais | [celsofranca@dcc.ufmg.br](mailto:celsofranca@dcc.ufmg.br) ]  
Marcos Carvalho   [ Universidade Federal de Minas Gerais | [marcoscarvalho@dcc.ufmg.br](mailto:marcoscarvalho@dcc.ufmg.br) ]  
Marcelo Ganem   [ Universidade Federal de Minas Gerais | [marceloganem@dcc.ufmg.br](mailto:marceloganem@dcc.ufmg.br) ]  
Gabriel Teixeira   [ Universidade Federal de Minas Gerais | [gabrielmedeiros@dcc.ufmg.br](mailto:gabrielmedeiros@dcc.ufmg.br) ]  
Gabriel Jallais   [ Universidade Federal de Minas Gerais | [gabrieljallais@dcc.ufmg.br](mailto:gabrieljallais@dcc.ufmg.br) ]  
Alberto H. F. Laender   [ Universidade Federal de Minas Gerais | [laender@dcc.ufmg.br](mailto:laender@dcc.ufmg.br) ]  
Marcos A. Gonçalves   [ Universidade Federal de Minas Gerais | [mgoncalv@dcc.ufmg.br](mailto:mgoncalv@dcc.ufmg.br) ]

 Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627 - Pampulha, Belo Horizonte, MG, 31270-901, Brazil.

Received: 25 February 2023 • Published: 20 October 2023

## 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 (contextual) 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 only by digits and punctuation, such as IDs and phone numbers, as well as long composed names. In this article, 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. To provide further context to the training process of NER+RE, we propose a data augmentation technique based on Generative Pretrained Transformers (GPT). We evaluate our strategies using real data from public administration documents (official gazettes and biddings) and court lawsuits. Our results show that our pre- and post-processing strategies, when used co-jointly, allows significant improvements on NER+ER effectiveness, while we also show the benefits of using GPT for training data augmentation.

**Keywords:** Named Entity Recognition, Relation Extraction, Contextual Embeddings, Contextual Reinforcement, Entity Delimitation, Training Data Augmentation, GPT, Public Administration

## 1 Introduction

Named Entity Recognition (NER) and Relation Extraction (RE) tasks are useful in various data management applications such as record deduplication [Silva *et al.*, 2019], data integration [Brunner and Stockinger, 2020], knowledge bases construction [Niu *et al.*, 2012], and information retrieval (IR) [Caputo *et al.*, 2009]. In the context of these applications, while the NER task aims to identify entities mentioned in a text (e.g., names of people and organizations), as well as classify them into a predefined set of categories, the RE task seeks to identify and classify the possible existing relationships between the entities of the text (e.g., the employee-employer relationship existing between a person and an organization) [Eberts and Ulges, 2020].

In this scenario, neural architectures based on transformers (e.g., Bidirectional Encoders from Transformers (BERT) [Devlin *et al.*, 2019] constitute the state-of-the-art in NER and RE. Among these methods, a recent representative approach is Span-based Entity and Relation Transformer (SpERT) [Eberts and Ulges, 2020], which jointly performs the two aforementioned tasks of NER and RE. SpERT, as

well as other methods that use transformers, is based on semantic (i.e., contextual) aspects of word sequences, which may be absent, for example, in entities composed only of digits and punctuation marks (e.g., ID and phone numbers). On the other hand, these entity types have well-behaved patterns that can be captured via regular expressions, which can enhance and reinforce the semantic context of sentences in which such patterns occur.

Thus, the main objective of this article is to propose new techniques that enhance the semantic context used to recognize entities and relationships. Towards this direction, we propose a new pre-processing step for the input text. This step employs regular expressions to highlight in a text specific entities such as CPF<sup>1</sup> (the main identification number for Brazilian residents) and CNPJ<sup>2</sup> (the main identification number for Brazilian organizations). It boosts the recognition not only of these more regular types of entity, but also of other ones that occur next to them or in the same context – for example, a CNPJ often occurs next

<sup>1</sup>Cadastro de Pessoas Físicas

<sup>2</sup>Cadastro Nacional de Pessoas Jurídicas

to the name of the organization associated with it.

Despite the potential improvements of this semantic reinforcement technique, we have empirically verified that the original SpERT can return different parts of a single entity as if they were different entities, e.g., “João da Silva” and just “João” when name and last name appear together. It may also include words that are not part of the entity (e.g., in “STATE SECRETARY REGISTERED”, the method erroneously included the expression “REGISTERED” as part of the entity, probably due to the capitalization pattern of the text). To better delimit the entities, we also propose a post-processing step, aiming at unifying entities that were not well delimited by SpERT and similar methods. This step chooses the most likely mention for each set of mentions that are overlapping in the original method’s output. The probability is given by the SpERT output itself, which associates a score with each recognized entity.

One issue with the current state-of-the-art NER+RE methods based on Transformers is that they require a lot of data to become fully effective. The use of active learning helps to obtain high quality training data, but it is still difficult and expensive to obtain large amounts of training data through human expert annotations. Accordingly, to expand the training data set in these collections and improve the training of the SpERT method, we proposed a synthetic training data generation technique based on Generative Pre-training Transformers (GPT).

GPT-based language models [Brown *et al.*, 2020] have shown great potential for several Natural Language Processing (NLP) tasks, requiring only a few labeled examples for few-shot learning. However, the very-large-scale nature of these models requires a high-cost infra-structure. Additionally, the most successful models, such as GPT-3 and ChatGPT, are not open source and are only accessible through APIs provided by companies such as OpenAI<sup>3</sup> and HuggingFace<sup>4</sup>. In this context, privacy issues arise as many official documents contain sensitive information that cannot be submitted through APIs. Our method generates training data using public data only. During the inference stage, smaller models such as SpERT can be used instead, along with lower-cost infra-structures.

We evaluate our new techniques using three datasets of different document types: court lawsuits, official diaries, and bidding documents. Data from court lawsuits were obtained from the LENER-BR [Luz de Araujo *et al.*, 2018] collection, while data from official diaries were collected from the Official Gazette of the State of Minas Gerais (OG-MG), comprising 208 documents from the official gazette published during the 2016 year. Finally, the bidding documents consist of previously documents collected from webpages of cities of the Association of the City Halls of Minas Gerais (*Associação dos Municípios Mineiros – B-AMM*). Using predefined sets of 10-13 entity labels and 9-10 relationship labels, we manually labeled representative samples of OG-MG and B-AMM data, selected using active learning techniques [Wang *et al.*, 2021].

Our experimental results show that the pre-processing

strategy leads to gains of 4% up to 9% in recall in the NER and RE tasks, while the post-processing step is responsible for gains ranging from 16% up to 224% in precision in both NER and RE tasks, allowing more precise delimiting of entity mentions in the text. Both proposed strategies have a negligible additional cost (less than 2%) in relation to the total cost of recognizing entities and relationships. Moreover, our GPT-based training data augmentation strategy leads to large gains of up to 1064% in precision and 158% in recall with relation to using only manually annotated training data.

In short, the main contributions of this article are:

1. A semantic context enhancement strategy based on pre-processing of input data, generating results with larger coverage of entities and relationships (recall);
2. A strategy for delimiting entities in the text through post-processing of the results, increasing the precision in recognizing both entities and relations;
3. A data augmentation strategy based on GPT which allows significant gains in both precision and recall;
4. Generation of new collections of relevant data for evaluation and training of NER+RE algorithms in official documents, publicly available<sup>5</sup>.

This article extends our previous work [Belém *et al.*, 2022] including a new evaluation dataset (B-AMM) and the new GPT-based data augmentation technique.

It is worth mentioning that the strategies proposed in this article are being used for final practical government applications to support the Public Ministry of Minas Gerais. Among them, we can mention a search tool in official documents and a semantic classification tool.

The remainder of this article is organized as follows. Section 2 presents related work, while Section 3 presents the problem statement. Section 4 describes the proposed strategies, while Section 5 describes the evaluation methodology. Section 6 presents experimental results and discussion, while Section 7 concludes the article and points out directions for future work.

## 2 Related Work

Several studies have addressed the NER+RE task with deep learning architectures. Next, we present an overview of these related works grouped into discriminative (Section 2.1) and generative approaches (Section 2.2).

### 2.1 Discriminative Approaches

NER and RE discriminative approaches classify each token of the text (token-based approaches) or each sequence of tokens (span-based approaches) into one entity type. Token-based methods additionally identify whether the word belongs to the beginning, middle, or end of the identified entity [Finkel *et al.*, 2005; Patil *et al.*, 2020], and latter join consecutive tokens that were classified as the same entity type. Span-based methods first enumerate all spans (sequences of

<sup>3</sup><https://openai.com>

<sup>4</sup><https://huggingface.co/>

<sup>5</sup><https://github.com/MPMG-DCC-UFMG/M01>

tokens) that are smaller than a given threshold and then classify each of the enumerated spans [Eberts and Ulges, 2020; Fu et al., 2021; Liu et al., 2021]. Additionally, for each pair of spans with a high probability of being an entity, these techniques infer: (i) whether there is a relationship between them and (ii) the type of relationship.

Examples of token-based methods traditionally used in NER include those based on Conditional Random Fields (CRF's) [Finkel et al., 2005; Patil et al., 2020]. CRF's are probabilistic models that infer the category of each  $t$  token from a text by exploring attributes of  $t$  (e.g., uppercase and lowercase letter patterns) and attributes of tokens adjacent to  $t$ , as well as the inferences made about them.

Span-based strategies [Fu et al., 2021] have received recent attention due to their excellent results, which are usually better than token-based approaches. They are also easily modeled using neural architectures based on transformers, using only “raw” attributes (word embeddings), without the need to elaborate and extract complex features from the data. In addition, such strategies allow the extraction of entities that overlap in the text. For example, in the case of “Ministério Público de Minas Gerais”, two entities could be considered: “Ministério Público de Minas Gerais” and “Minas Gerais”. A disadvantage of this strategy, as our experiments show, is that it is often imprecise in delimiting an entity in the text. To deal with this problem, this article proposes to incorporate, into state-of-the-art techniques, a post-processing strategy for NER+ER results that makes the delimitation of entities more precise, without the need of additional training.

Neural networks based on transformers (e.g., Bidirectional Encoders from Transformers or BERT [Devlin et al., 2019]) represent the state-of-the-art in several natural language processing tasks Constantino et al. [2022], including the task of recognizing entities and their relationships (NER+ER). Among the NER+RE methods based on transformers, the Span-based Entity and Relation Transformer (SpERT) [Eberts and Ulges, 2020] stands out. SpERT encodes text spans into vector representations (embeddings) based on pre-trained models in order to classify them as one of a set of pre-defined categories of entities or eventually as a “non-entity”. Pairs of spans that the algorithm identifies as entities are also represented in the vector space and are eventually assigned to some predefined category of relation. Eberts & Ulges [2021] extended the SpERT architecture to include a grouping of mentions that refer to the same entity in different segments of a text.

Methods such as SpERT are strongly based on the semantic context of sequences of words in the text and may have difficulties in recognizing entities and relationships when there is little contextual information in such sequences (e.g., entities formed only by digits and punctuations, such as CPF's and telephone numbers). In this article, we circumvent this problem by means of a contextual reinforcement given by the marking of these entities through regular expressions that enhance the textual semantic context. This new step helps not only in identifying more regular entities, but also entities close to them. As a consequence, there is an improvement in the recognition of relationships between entities.

## 2.2 Generative Approaches

Generative approaches use statistical models to generate new data instances. The augmented data may leverage supervised learning by helping to fine-tune discriminative models. Huang et al. [2022b], for instance, proposed COPNER that produces augmented data, by employing prompt engineering, to increase the number of labeled samples for the NER task. It leverages contrastive learning with BERT to assign an entity to a sentence's tokens.

Similarly, Huang et al. [2022a] propose a prompt-based self-training two-step framework. In the first step, the authors employ a self-training approach with various label selection strategies to mitigate the error propagation of noisy pseudo-labels. In the second step, they fine-tune a BERT model using the high-confidence pseudo-labels and the original labels. In Wang et al. [2022], the authors leverage prompt learning to extract entity-related objects from images. They enrich the features by combining text and objects into a dense vector representation, which a multilayer perceptron employs to predict entities.

Generally, state-of-the-art generative approaches typically utilize prompt learning to carry out a pseudo-labeling procedure on unlabeled data. In contrast, our approach goes beyond by not only generating labels but also producing the text for new training instances. Furthermore, while existing methods primarily concentrate on entity recognition, which is only a portion of the problem discussed in this article, we tackle the more complex task of identifying not only entities but also their relationships.

## 3 Task Description

In this section we describe the Named Entity Recognition (NER) and Relation Extraction (RE) tasks. The former seeks to extract and classify entities mentioned in texts, allowing their separation into predefined categories such as: person, local and organizations, monetary values, CPF's, CNPJ's, telephone numbers, among others.

Typically, the NER task processes an unstructured text, such as the following illustrative example: “The company XYZ Ltda., inscribed in CNPJ 12.345.678/1234-56, is located at 1001, A Street”, producing a block of annotated text highlighting the named entities and their respective categories, with the tags “ORGANIZATION”, “CNPJ” and “ADDRESS” as shown in Figure 1.

In contrast, the RE task produces a list of relations between the entities found in the text, represented as arrows in Figure 1. Formally, a relation can be defined as an  $(e_1, r, e_2)$  triple, where  $e_1$  and  $e_2$  are entities and  $r$  is the type of the relation between them. In this article, we tackle the combined task of NER+RE, aiming to simultaneously recognize entities and their associated relations.

## 4 Proposed Strategies

We propose extensions of NER+RE strategies that include: (i) a pre-processing technique of the input text, which we call Contextual Reinforcement (ConRe); (ii) a post-processing

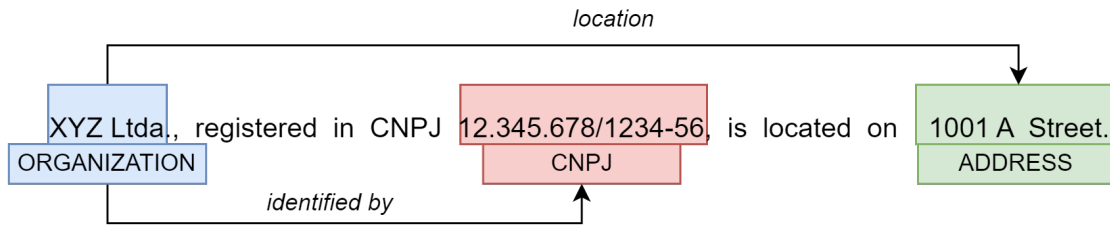


Figure 1. Example of an NER+RE assignment.

strategy of the result, called Entity Delimitation (EntD); and (iii) a data augmentation strategy based on generative language models. The ConRe technique (Subsection 4.1) can be applied to any dataset that contains at least one regular entity type, i.e., entities that can be extracted using regular expressions. The EntD strategy (Subsection 4.2) can be applied to any method that produces an estimate of confidence in the classification of entities. As these scenarios are quite common in NER+RE [Zhang *et al.*, 2018], such strategies are widely applicable. Finally, our data augmentation approach (Section 4.3) is general for any NER+RE scenario. In this article, we use the state-of-the-art Span-based Entity and Relation Transformer (SpERT) [Eberts and Ulges, 2020] as the base NER+RE strategy to be extended.

#### 4.1 Contextual Reinforcement (ConRe)

Our ConRe strategy consists of inserting preliminary markings in the text that indicate the type, beginning, and end of some of the entities listed in the text. To make this possible, we focus on entities that have a more “well-behaved” pattern that can be easily captured by regular expressions.

In the particular case of the collection of official gazettes, we identified several entity types with regular characteristics. One of them is CPF, which is composed by 11 digits, eventually interspersed with periods and a hyphen that separates the two verifying digits from the other digits. This can be described by the following regular expression:  $\backslash d\{3\}\.?\backslash d\{3\}\.?\backslash d\{3\}\-?\backslash d\{2\}$ . This and other useful regular expressions for identifying entities in official documents can be found on the repository page (see link in Section 1).

Thus, for each sequence of characters in the text that matches one of the regular expressions, the expression  $[type]$  is inserted, immediately before and after the identified sequence, where  $type$  is a name corresponding to the type of entity captured by the regular expression (e.g., CPF, DATE). Both the data that are provided as training to the base strategy and the new data from which entities and relationships are to be extracted are marked. After extraction, we remove the tags for presentation purposes.

An alternative to tags would be to recognize regular entities directly with regular expressions, without using the base strategy. We verified that the markings also helped SpERT in the identification of other entities close to the regular ones, as it provides more context for this method.

#### 4.2 Entity Delimitation (EntD)

Another challenge in the NER+RE task is the typical difficulty in delimiting some entity mentions. For instance, the algorithm may not capture a person’s full name if it includes many surnames. Dealing with this by varying the uppercase/lowercase letter pattern may not work in official documents, which often present the rest of the sentence in all caps, making it difficult to distinguish between proper names and common words. The same goes for organization names or even regular entities, depending on the spacing patterns and digit punctuation (e.g., “14/06” VS. “14 / 06” VS. “14-06”, etc).

In these cases, span-based approaches return different strings that refer to parts of the same entity (or even include words that are outside the boundaries of an entity). This guarantees the maintenance of the recall of mentions but damages the precision in the recognition of entities and, consequently, of relationships.

To mitigate this problem, we propose an Entity Delimitation (EntD) strategy that unifies the mentions referring to the same entity by using the confidence estimates of the base strategy itself (in this case, SpERT). To achieve this, for each set of entity mentions that overlap in the algorithm’s output, we choose only the most likely one according to the algorithm’s estimate.

Figure 2 illustrates the proposed steps for the NER+RE task, presenting an example and the expected results for it. In step (a), ConRe identified a sequence of characters that matches the regular expression for CNPJ’s, marking this sequence, thus facilitating the subsequent identification of this entity by any NER strategy (base method) in step (b). The result of the base method found two overlapping entities: “XYZ” and “XYZ Ltda.”, with 85% and 90% chance, respectively, of being an ORGANIZATION. In the final step (c), the EntD eliminates this “duplicity”, keeping only the most likely instance.

#### 4.3 GPT-based Data Augmentation

Our third strategy for providing and reinforcing the context, required for state-of-the-art NER+RE methods, is to expand the training dataset. These methods usually require large amounts of labeled data, which may be difficult or expensive to obtain using human expert annotation.

Generative Pre-training Transformers (GPT), which are considered Large Language Models (LLM), have shown great potential for NLP tasks, requiring only a few labeled examples for few-shot learning. The most powerful GPT models have structures that are orders of magnitude larger

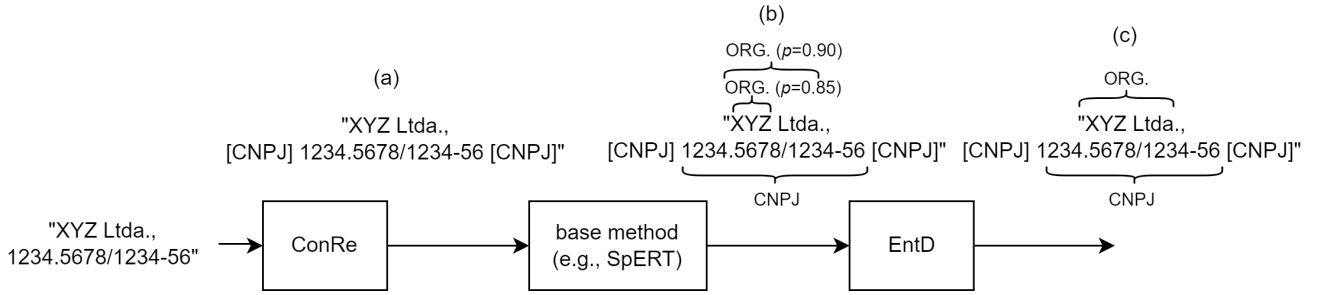


Figure 2. Steps and example of the proposed NER+RE processing strategy.

than previous Transformers, allowing them to produce texts similar to those written by humans.

However, to run these large-scale models, high computational power is required, usually in the form of multiple GPU's, which is expensive to acquire and maintain. Additionally, the most successful models, such as GPT-3 and ChatGPT, are not open source and are only accessible through API's provided by companies such as OpenAI and HuggingFace.

In this work, we leverage large GPT's to generate and label new training data for NER+RE methods. Directly using GPT's for NER+RE tasks is not feasible in our scenario due to privacy and cost concerns. Privacy issues arise as many official documents contain sensitive information that cannot be submitted through API's. Cost issues stem from the fact that the best GPT models are not free and the cost is proportional to the amount and size of the requests.

Our method generates training data using only public data and submits requests only once (before the offline training stage), reducing the cost. During the inference stage, smaller models such as SpERT (after the training with our augmented data) can be used instead, keeping the process cost-effective and private within our own (small) infrastructure.

Specifically in the case of our data augmentation strategy, we build one GPT prompt for each relation type as follows. Given a set of relations  $\{r_1, r_2, \dots, r_n\}$  present in our initial training dataset, for each relation  $r_i$  consisting of a pair of entities  $(E_{r_i}, E'_{r_i})$ , the prompt has the following structure:

```
[Text]:  $E_{r_1} \text{ context}_{r_1} E'_{r_1}$ 
[type( $E_{r_1}$ )]:  $E_{r_1}$ 
[type( $E'_{r_1}$ )]:  $E'_{r_1}$ 
####
...
[Text]:  $E_{r_n} \text{ context}_{r_n} E'_{r_n}$ 
[type( $E_{r_n}$ )]:  $E_{r_n}$ 
[type( $E'_{r_n}$ )]:  $E'_{r_n}$ 
####
[Text]:
```

where  $\text{context}_{r_i}$  is the text that occurs between  $E_{r_i}$  and  $E'_{r_i}$ ,  $\text{type}(E)$  is the label of the entity  $E$ , and  $\text{###}$  is the sentence separator. The final [Text] indicates how GPT must start its text generation process, producing new examples following the patterns established by the previous examples. Figure 3 illustrates an example of our described prompt.

To increase diversity in the generation of an annotated text, examples are supplied in a random order. If all the examples cannot be accommodated within one prompt, which is typi-

Figure 3. Example of GPT prompt for generation of annotated sentences expressing localization relations.

```
[Text]: XYZ Inc. is located at 1001 A Street.
[Organization]: XYZ Inc.
[Address]: 1001 A Street.
####
[Text]: State Health Department.
Address: 3045 E Avenue, Route 49, Oswego.
[Organization]: State Health Department
[Address]: 3045 E Avenue, Route 49, Oswego
####
[Text]:
```

cally limited to 1000-4000 characters, the ones that will be included in the prompt are randomly selected, for each text generation request.

Another way of controlling variability is to set the tuning parameter *temperature*, choosing one value between 0 (lowest variability) and 1 (highest variability). The temperature parameter controls the level of randomness or creativity in the generated text. It is a scaling factor applied to the logits (output of the final layer of the neural network) before they are converted into probabilities through the *softmax* function. A higher temperature value results in a more diverse and varied output, as the model is more likely to select less probable options. On the other hand, a lower temperature value leads to more conservative and predictable outputs, as the model is more likely to select the most probable option [Brown et al., 2020].

## 5 Evaluation Methodology

In this section, we describe the data collections (Subsection 5.1) and the metrics used in the experimental evaluation (Subsection 5.2,) as well as the SpERT parameterization (Subsection 5.3).

### 5.1 Data Collections

To evaluate the effectiveness of the proposed strategies, we employed three real data collections: the court lawsuits dataset LENER-BR [Luz de Araujo et al., 2018], the 2016b Official Gazette of the State of Minas Gerais (OG-MG), in which we labeled a representative sample of 214 sentences, and the Bidding Documents of the Associação Mineira dos Municípios (B-AMM), in which we labeled a sample of 78 sentences.

The OG-MG sample of 214 representative and diverse sentences was selected through an active learning strat-

**Table 1.** Data collections

	LENER-BR	OG-MG	B-AMM
Text type	Court lawsuits	Official diaries	Bidding documents
#sentences with labels	10000	214	78
#entity mentions	9800	1255	782
#relations	-	817	234
#entity types	6	10	13
#relation types	-	10	9

egy [Wang *et al.*, 2021]. Based on this sample, 10 relevant types of entity and 10 types of relationship present in official gazettes were identified. The types of entity identified were: PERSON, ORGANIZATION, DATE, LOCATION, COMPETENCE (public administration positions), LEGISLATION, NUMBER\_ACT (identifiers of acts performed by the public administration), CPF, CNPJ, and MASP (a public servant registration number). More details about these entity types, as well as the types of relationship between them can be found in the available data repository.

Similarly, for the B-AMM dataset we identified the following entity types: PERSON, ORGANIZATION, DATE, LOCATION, COMPETENCE, PROCESS, CONTRACT, BIDDING\_MODALITY, CPF, CNPJ, MONETARY\_VALUE, LEGISLATION and PHONE\_NUMBER.

Each instance (i.e., sentence) of the selected sample was inspected by at least three members of our group (Computer Science undergraduate and graduate students at UFMG) who were properly instructed on how the labeling should be done. As the gold standard for each labeled entity and relation, the most voted label prevailed. There was disagreement in less than 10% of both, entity and relation annotations.

Finally, due to the high annotation costs and the need for a larger amount of training data, we expanded the OG-MG and B-AMM datasets using our proposed GPT-based data augmentation strategy described in Section 4.3. Our code, as well as the synthetic and real data collections, are available to ensure reproducibility of results and to contribute to the future development and evaluation of new NER and RE strategies (see link in Section 1).

## 5.2 Evaluation Metrics

We assess the experimental results using Precision, Recall, and F1 metrics, which capture different effectiveness aspects of the NER-ER task. Considering  $x$  as a type of entity or relation (e.g.,  $x = \text{PERSON}$  or  $x = \text{located\_at}$ ), the Precision of the algorithm to recognize entities or relations of type  $x$  is calculated as:

$$\text{Precision}(x) = \frac{\text{Correct}}{\text{Total times the algorithm recognized type } x}$$

Recall, on the other hand, measures how much the algorithm is able to cover *all* mentions to entities of type  $x$  (or relations of type  $x$ ) in a given text:

$$\text{Recall}(x) = \frac{\text{Correct}}{\text{Total times type } x \text{ is mentioned in the text}}$$

Finally,  $F1(x)$  is defined as the harmonic mean between  $\text{Precision}(x)$  and  $\text{Recall}(x)$ .  $F1(x)$  penalizes the value of the metric if any of the two component metrics (or both) is low.

To aggregate the effectiveness measures for all entity and relationship types, micro and macro averages are extracted for each metric. The Macro version consists of the average of measures over all entity/relationship types. The Micro version considers the overall effectiveness of the algorithm independently of the effectiveness by entity/relationship type. While the macro average considers all entity (or relation) types equally important, independently of their frequency in the dataset, the micro average tends to give more importance to more frequent entity/relation types.

For purposes of evaluating the proposed strategies, the data collections are divided into three partitions, called training, validation, and test sets. Training is used to learn the NER+RE model, while validation is used to adjust parameters. Finally, the results are reported on the test set. For the LENER-BR collection, we exploited the same training, validation, and test sets provided by the authors. For the OG-MG and B-AMM collections, the test set is formed by real data (half of the sentences manually labeled), while the validation and training sets are formed by real (the other half of the labeled data) and synthetic data generated by our augmentation strategy presented in Section 4.3. The results of data augmentation process are discussed in Subsection 6.1.

As language model for our data augmentation process, we exploited the GPT model BLOOM<sup>6</sup>. It has the same number of model parameters as GPT-3 and allows a relatively large number of requests for free. In terms of the amount of synthetic data, we generate approximately 2000 new instances for both: the OG-MG and B-AMM datasets. To avoid duplicated instances, or very-similar instances, which could bias the model towards them, we removed generated instances containing the same pair of named entities or the same context (text between the pair of entities). After the removal of redundant instances, each dataset remained with about 400 instances, which at first sight may be considered a small amount of data but, as we empirically show, is enough to produce large effectiveness gains in the NER+RE task.

## 5.3 Parameterization

For the parameterization of the SpERT strategy, we used the values recommended by its authors [Eberts and Ulges, 2020], namely: learning rate  $l_r = 5 \times 10^{-5}$ , number of training epochs  $t = 20$ , number of negative examples per sentence  $n^- = 100$  (both for entities and relations) and size of each batch  $b_s = 2$ .

<sup>6</sup><https://huggingface.co/bigscience/bloom>

For the training data augmentation, we set the temperature parameter of the text generation process as 0.7, as it is widely used and reflects a good tradeoff between variability and accuracy.

## 6 Experimental Results

In this section, we first show the results of the training data augmentation for OG-MG and B-AMM datasets using GPT-generated synthetic phrases (Subsection 6.1). Next, we present the results of the proposed pre-processing (ConRe) and post-processing (EntD) strategies (Subsection 6.2).

### 6.1 GPT-Based Expansion of Training Data with Synthetic Phrases

Table 2 shows the results of the macro version of Precision, Recall, and F1 obtained by the SpERT strategy in the OG-MG and B-AMM collections with and without training data augmentation. The results for the micro averages are similar; we omitted them for the sake of space. Best results as well as statistical ties according to a two-sided t-test with  $\rho < 0.05$  are highlighted in bold.

We note that the addition of synthetic training data allowed gains ranging from 21% up to 183% in precision and 9% up to 115% in F1 in entity recognition (NER), when compared to the use of only manually-labeled training data. For the relation extraction (RE) task, it achieved even higher gains: 81% up to 1024% in precision and 68% up to 856% in F1. The only case with a small loss (in Recall) occurred in the recognition of entities of the OG-MG dataset, which may be explained by the possible introduction of some noise by the synthetic data.

Examples of noise in the synthetic data include the generation of (i) texts that deviate from the requested relation type, (ii) non-labeled entities that should be labeled, and (iii) wrongly labeled entities. However, we observed a relatively low fraction of labeling mistakes (about 5%) in a small sample we manually inspected, which reveals the power of large GPT models. This noise amount is not necessarily higher than human annotation errors [Zhu *et al.*, 2022]. This, added to the fact that our manually labeled sample was very small, specially for the B-AMM dataset, explains why there are no losses in any evaluation metric for this dataset. Thus, it is better to provide a more complete training dataset, even if it is synthetically generated, than being limited to a very small manually labeled dataset, when considering state-of-the-art NER+RE techniques based on small to medium scale neural networks such as SpERT.

On the other hand, the large obtained gains are explained by the scarceness of real labeled data, which is even more crucial in the case of the (small) B-AMM dataset. The difference of results also occur because SpERT, as a supervised Transformer method, depends on a significant amount of labeled training examples to be effective. However, obtaining actual labeled data samples by manually labeling in this task (NER+ER) and domain (legal documents) is hard and costly, especially when it is also necessary to identify relations.

To name a few obstacles in NER+RE annotation, we cite: (i) the process is naturally laborious: in a single sentence or paragraph, there are usually multiple entities and relations to identify; (ii) it is often necessary the help of domain experts to correctly identify entities and relations, and (iii) even using an annotation tool such as Webanno [Eckart de Castilho *et al.*, 2016], the typically large number of entities (highlighted in the text) and relations (arrows from some entities to others) make the annotation process visually confusing.

To summarize, our results show the benefits of automatically expanding the training with a synthetic phrase generation strategy, which requires only a small labeled dataset as seed. Our strategy exploits the knowledge embedded in large language models such as GPT to automatically generate new labeled training instances.

### 6.2 Effectiveness of Proposed Strategies

We present the effectiveness results of our proposed pre-processing (ConRe) and post-processing (EntD) techniques. Tables 3 and 4 show macro and micro results, respectively, for the precision, recall, and F1 metrics for the OG-MG data collection, while Tables 3 and 4 show the corresponding results for the B-AMM dataset. Finally, Table 7 shows these measures for entity recognition in the LENER-BR collection. For each table, the best results (and statistical ties according to a two-sided t-test with  $\rho < 0.05$ ) are highlighted in bold.

Note that both ConRe and EntD strategies were applied to the OG-MG and B-AMM collections, while only EntD was applied to the LENER-BR. This happened because the latter does not contain labels for regular entities, such as CPFs and telephone numbers, which do not require the semantic enrichment provided by ConRe.

Next, we discuss the results referring to the isolated application of ConRe and EntD (Subsections 6.2.1 and 6.2.2 respectively), as well as the results of the joint application of both techniques (Subsection 6.2.3). For all those analysis, we exploited the expanded version of the training datasets.

#### 6.2.1 Effectiveness of ConRe

Comparing the results of the first and third lines of Tables 3 and 4, we can observe that ConRe, when applied in isolation (without the EntD post-processing step), generates modest gains, ranging from 4% up to 9% in recall (for both macro and micro averages, in both tasks: NER and RE). It also produces precision gains of 12% up to 30% in the B-AMM dataset; no precision gains were observed in the OG-MG dataset. As we shall see (Section 6.2.3), when EntD and ConRe are exploited co-jointly, we observe gains in all scenarios (all evaluation metrics and datasets).

Such gains occur because ConRe induces the semantic enrichment of sentences, which helps SpERT to identify types of entities such as CPFs and CNPJs, which are basically composed of numbers, with little associated semantics. In these cases, using regular expressions tends to be a more effective solution. Indeed, often, acronyms such as CPF and CNPJ already accompany the numbers corresponding to these identifiers in the text. Despite this, in some cases SpERT had difficulty delimiting these numbers, due to variations in spaces

**Table 2.** Precision, Recall, and F1 macro values (and 95% confidence intervals) with real training data only and with training data expanded with synthetic sentences. Best scores (and statistical ties) are in bold.

Dataset	Training data	Entities (NER)			Relations (RE)		
		Precision	Recall	F1	Precision	Recall	F1
OG-MG	Only Real	0.570 ± 0.027	<b>0.857 ± 0.019</b>	0.671 ± 0.026	0.325 ± 0.032	0.657 ± 0.033	0.392 ± 0.033
	Augmented	<b>0.688 ± 0.026</b>	0.804 ± 0.022	<b>0.734 ± 0.024</b>	<b>0.588 ± 0.034</b>	<b>0.755 ± 0.029</b>	<b>0.657 ± 0.033</b>
B-AMM	Only Real	0.207 ± 0.038	0.729 ± 0.042	0.307 ± 0.043	0.020 ± 0.022	0.272 ± 0.072	0.036 ± 0.030
	Augmented	<b>0.587 ± 0.046</b>	<b>0.813 ± 0.036</b>	<b>0.661 ± 0.044</b>	<b>0.231 ± 0.068</b>	<b>0.701 ± 0.074</b>	<b>0.341 ± 0.076</b>

**Table 3.** Macro-Precision, Macro-Recall, and Macro-F1 (and 95% confidence intervals) with and without the application of ConRe and EntD techniques, in the OG-MG collection. Best scores (and statistical ties) are in bold.

ConRe?	EntD?	Entities (NER)			Relations (RE)		
		Precision	Recall	F1	Precision	Recall	F1
No	No	0.688 ± 0.026	0.804 ± 0.022	0.734 ± 0.024	0.588 ± 0.034	0.755 ± 0.029	0.657 ± 0.033
No	Yes	0.796 ± 0.022	0.774 ± 0.023	0.781 ± 0.023	<b>0.733 ± 0.030</b>	0.714 ± 0.031	0.709 ± 0.031
Yes	No	0.672 ± 0.026	<b>0.838 ± 0.020</b>	0.742 ± 0.024	0.518 ± 0.034	<b>0.786 ± 0.028</b>	0.616 ± 0.033
Yes	Yes	<b>0.820 ± 0.021</b>	<b>0.823 ± 0.021</b>	<b>0.820 ± 0.021</b>	<b>0.726 ± 0.031</b>	<b>0.770 ± 0.029</b>	<b>0.741 ± 0.030</b>

and punctuation between them. Regular expressions helped extract more instances of mentions of these entity types, and post-processing, as will be discussed later, helped to narrow them down more precisely.

Even in cases of more irregular entities (for example, names of people), there have been significant improvements, as they often occur next to regular entities. For instance, a CPF often occurs next to the name of the person associated with it, while a CNPJ tends to occur next to the name of the associated organization.

### 6.2.2 Effectiveness of EntD

When applied in isolation, EntD produces macro-precision gains ranging from 16% up to 38% for NER and corresponding gains ranging from 25% up to 224% for RE. Similar gains are observed for micro-precision. As expected, EntD does not increase recall, because it only eliminates improbable mentions that overlap other mentions in the output. As a consequence of the relatively high gains in precision, with negligible losses in recall, there are also gains in Macro-F1 (6%-18% for NER and 8%-103% for RE) and in Micro-F1 (7%-20% for NER and 14%-107% for RE).

The highest gains were observed in the B-AMM collection. In the LENER-BR, EntD promotes modest but statistically significant gains of up to 2% in macro and micro-precision. This is due to the lower complexity of the LENER-BR collection, which has only six entity types and no labeled relationships. For more complex collections such as OG-MG and B-AMM, the proposed strategies have great potential to improve results.

### 6.2.3 Effectiveness and Efficiency of the Joint Application of the ConRe and EntD Strategies

Comparing the results of the joint application of the ConRe and EntD strategies (fourth line of Tables 3-6) with the result of the original SpERT strategy (first line of the same tables), we obtain very high gains in precision (ranging from 19% to 55% for NER, and from 24% up to 243% for RE), without (statistically) harming recall (in case of B-AMM), or even (statistically) improving it (in case of OG-MG) by smaller

margins. This result is due to the increases in recall generated by ConRe and the high improvements in precision produced by EntD, as discussed earlier. ConRe allows for greater coverage of mentions of entities (and consequently of relationships), although it may generate overlapping results, in terms of inaccuracies in the delimitation of entities, a problem that exists with or without pre-processing. This problem, in turn, is tackled by applying the EntD post-processing technique.

Regarding the efficiency of the proposed strategies, it is worth mentioning that their additional cost with relation to the original SpERT strategy is negligible. The total ConRe response time for all 214 OG-MG test sentences was 0.04s, which is equivalent to an increase of 1.3% of the total SpERT runtime, which was 3.06s. EntD response time, also negligible, accounts for less than 1% of total SpERT's execution time. Such results were obtained on a 6-core AMD Ryzen 5 5600X processor with 64GB of RAM.

## 7 Conclusions and Future Work

In this article, we proposed Contextual Reinforcement (ConRe), Entity Delimitation (EntD) and GPT-Based Training Data Augmentation strategies for the Entity Recognition and Relation Extraction (NER+RE) task. Such strategies are broadly applicable to many types of unstructured data. Among these data, official documents stand out, as they usually present a rich set of entities that can be recognized by regular expressions.

The ConRe strategy performs a preliminary marking of regular entities in the text, allowing the identification of entities that even a state-of-the-art algorithm such as SpERT may not be able to capture without this type of pre-processing. The EntD strategy performs post-processing of the result in order to unify overlapping entity mentions in the output, increasing the precision of the recognition. Their joint application allowed to delimit entities and relations up to 224% better when compared to the state-of-the-art strategy. Both strategies have a negligible additional cost in relation to the total cost of the task. Moreover, we showed the benefits of our training data augmentations using GPTs, which can successfully generate new annotated training



**Table 4.** Micro-Precision, Micro-Recall and Micro-F1 (and 95% confidence intervals) with and without the application of pre- and post-processing techniques, in the OG-MG collection. Best scores (and statistical ties) are in bold.

ConRe?	EntD?	Entities (NER)			Relations (RE)		
		Precision	Recall	F1	Precision	Recall	F1
Not	Not	0.705 ± 0.025	0.858 ± 0.019	0.774 ± 0.023	0.597 ± 0.034	0.788 ± 0.028	0.680 ± 0.032
Not	Yes	0.827 ± 0.021	0.831 ± 0.021	0.829 ± 0.021	<b>0.789 ± 0.028</b>	0.759 ± 0.029	0.774 ± 0.029
Yes	Not	0.720 ± 0.025	<b>0.892 ± 0.017</b>	0.797 ± 0.022	0.536 ± 0.034	<b>0.819 ± 0.026</b>	0.648 ± 0.033
Yes	Yes	<b>0.856 ± 0.019</b>	<b>0.872 ± 0.019</b>	<b>0.864 ± 0.019</b>	<b>0.797 ± 0.028</b>	<b>0.804 ± 0.027</b>	<b>0.801 ± 0.027</b>

**Table 5.** Macro-Precision, Macro-Recall, and Macro-F1 (and 95% confidence intervals) with and without the application of ConRe and EntD techniques, in the B-AMM collection. Best scores (and statistical ties) are in bold.

ConRe?	EntD?	Entities (NER)			Relations (RE)		
		Precision	Recall	F1	Precision	Recall	F1
No	No	0.587 ± 0.046	0.813 ± 0.036	0.661 ± 0.044	0.231 ± 0.068	0.701 ± 0.074	0.341 ± 0.076
No	Yes	0.807 ± 0.037	0.754 ± 0.040	0.778 ± 0.039	0.749 ± 0.070	0.663 ± 0.076	<b>0.693 ± 0.074</b>
Yes	No	0.683 ± 0.043	<b>0.889 ± 0.029</b>	0.744 ± 0.041	0.308 ± 0.074	<b>0.737 ± 0.071</b>	0.414 ± 0.079
Yes	Yes	<b>0.882 ± 0.030</b>	0.777 ± 0.039	<b>0.821 ± 0.036</b>	<b>0.782 ± 0.067</b>	0.643 ± 0.077	<b>0.691 ± 0.074</b>

instances, allowing to train NER+RE models with a low labeling effort; we just need a small (high quality) training seed, which can be obtained with active learning.

In future work, we intend to evaluate the effectiveness of the proposed strategies when applied to other NER+RE methods (not only SpERT), as well as to other data collections and tasks. We also intend to analyze in which situations it is possible to allow a higher level of overlapping entities (increasing recall) without harming precision. One idea could be by exploring co-occurrence of terms [Menezes *et al.*, 2010] among entities. Finally, we intend to study the role (and quality) of the initial training seed for the augmentation process and investigate whether it is possible, through better selective (active) sampling [de Freitas *et al.*, 2010; Cardoso *et al.*, 2017; Silva *et al.*, 2022], to produce synthetic data for training with even a smaller labeling effort.

## Acknowledgements

We would like to thank the Public Ministry of the State of Minas Gerais for its support under the *Capacidades Analíticas* project.

## Funding

Work partially supported by the authors' individual grants from CAPES, CNPq and FAPEMIG.

## Authors' Contributions

Fabiano Belém, Cláudio Valiense and Celso França: Conceptualization, Writing (review & editing), Methodology, Validation. Marcos Carvalho, Marcelo Ganem, Gabriel Teixeira and Gabriel Jalais: Writing (review), Methodology, Validation. Alberto H. F. Laender and Marcos A. Gonçalves: Conceptualization, Writing (review & editing), Validation, Project Management, Supervision.

## Competing interests

The authors declare that they have no competing interests.

## Availability of data and materials

The datasets and the source code of the current study are available at <https://github.com/MPMG-DCC-UFGM/M01>.

## References

- Belém, F. M., Ganem, M., Celso França, M. C., Laender, A. H. F., and Gonçalves, M. A. (2022). Reforço e Delimitação Contextual para Reconhecimento de Entidades e Relações em Documentos Oficiais. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 292–303, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2022.224650.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Brunner, U. and 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., and Semeraro, G. (2009). Boosting a Semantic Search Engine by Named Entities. In *Foundations of Intelligent Systems*, pages 241–250.
- Cardoso, T. N. C., Silva, R. M., Canuto, S. D., Moro, M. M., and Gonçalves, M. A. (2017). Ranked batch-mode active learning. *Inf. Sci.*, 379:313–337. DOI: 10.1016/j.ins.2016.10.037.
- Constantino, K., Cruz, V., Zucheratto, O., França, C., Carvalho, M., Silva, T. H. P., Laender, A. H. F., and Gonçalves, M. (2022). Segmentação e Classificação Semântica de Trechos de Diários Oficiais Usando Apre-

**Table 6.** Micro-Precision, Micro-Recall and Micro-F1 (and 95% confidence intervals) with and without the application of pre- and post-processing techniques, in the B-AMM collection. Best scores (and statistical ties) are in bold.

ConRe?	EntD?	Entities (NER)			Relations (RE)		
		Precision	Recall	F1	Precision	Recall	F1
No	No	0.558 ± 0.046	0.830 ± 0.035	0.667 ± 0.044	0.201 ± 0.065	<b>0.736 ± 0.071</b>	0.316 ± 0.075
No	Yes	0.818 ± 0.036	0.777 ± 0.039	0.797 ± 0.038	0.639 ± 0.077	0.669 ± 0.076	<b>0.653 ± 0.077</b>
Yes	No	0.624 ± 0.045	<b>0.884 ± 0.030</b>	0.732 ± 0.041	0.260 ± 0.071	<b>0.757 ± 0.069</b>	0.388 ± 0.078
Yes	Yes	<b>0.865 ± 0.032</b>	0.802 ± 0.037	<b>0.833 ± 0.035</b>	<b>0.691 ± 0.074</b>	0.649 ± 0.077	<b>0.669 ± 0.076</b>

**Table 7.** Micro and Macro Precision, Recall and F1 values (and 95% confidence intervals) with and without the application of EntD post-processing technique, in the LENER-BR collection. Best scores (and statistical ties) in bold.

EntD?	Micro Averages			Macro Averages		
	Precision	Recall	F1	Precision	Recall	F1
Not	0.834 ± 0.007	<b>0.863 ± 0.007</b>	0.847 ± 0.007	0.848 ± 0.007	<b>0.866 ± 0.007</b>	0.857 ± 0.007
Yes	<b>0.848 ± 0.007</b>	0.855 ± 0.007	<b>0.850 ± 0.007</b>	<b>0.863 ± 0.007</b>	0.858 ± 0.007	<b>0.861 ± 0.007</b>

- dizado Ativo. In *Anais do XXXVII Simpósio Brasileiro de Bancos de Dados*, pages 304–316, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2022.224656.
- de Freitas, J., Pappa, G. L., da Silva, A. S., Gonçalves, M. A., de Moura, E. S., Veloso, A., Laender, A. H. F., and de Carvalho, M. G. (2010). Active Learning Genetic Programming for Record Deduplication. In *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2010, Barcelona, Spain, 18-23 July 2010*, pages 1–8. IEEE. DOI: 10.1109/CEC.2010.5586104.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.
- Eberts, M. and 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. and 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.
- Eckart de Castilho, R., Mújdricza-Maydt, É., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 76–84, Osaka, Japan. The COLING 2016 Organizing Committee.
- Finkel, J. R., Grenager, T., and 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., and Liu, P. (2021). SpanNER: Named Entity Re-/Recognition as Span Prediction. In *Annual Meeting of the Association for Computational Linguistics*, pages 7183–7195.
- Huang, G., Zhong, J., Wang, C., Dai, Q., and Li, R. (2022a). Prompt-Based Self-training Framework for Few-Shot Named Entity Recognition. In Memmi, G., Yang, B., Kong, L., Zhang, T., and Qiu, M., editors, *Knowledge Science, Engineering and Management*, pages 91–103, Cham. Springer International Publishing.
- Huang, Y., He, K., Wang, Y., Zhang, X., Gong, T., Mao, R., and Li, C. (2022b). COPNER: Contrastive Learning with Prompt Guiding for Few-shot Named Entity Recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2515–2527, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Liu, C., Fan, H., and 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., and 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., and Bermejo, P. (2018). LeNER-BR: a Dataset for Named Entity Recognition in Brazilian Legal Text. In *International Conference on the Computational Processing of Portuguese (PROPOR)*, pages 313–323.
- Menezes, G. V., Almeida, J. M., Belém, F., Gonçalves, M. A., Lacerda, A., de Moura, E. S., Pappa, G. L., Veloso, A., and Ziviani, N. (2010). Demand-Driven Tag Recommendation. In Balcázar, J. L., Bonchi, F., Giannis, A., and Sebag, M., editors, *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part II*, volume 6322 of *Lecture Notes in Computer Science*, pages 402–417. Springer. DOI: 10.1007/978-3-642-15883-4\_26.
- Niu, F., Zhang, C., Ré, C., and Shavlik, J. W. (2012). Deep-Dive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. *VLDS*, 12:25–28.
- Patil, N., Patil, A., and 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., and Moro, M. (2019). Uma Análise Experimental do Impacto da Seleção de Atributos em Processos de Resolução de Entidades. In *SBBDD*, pages 37–48.

- Silva, R. M., Gomes, G. C. M., Alvim, M. S., and Gonçalves, M. A. (2022). How to build high quality L2R training data: Unsupervised compression-based selective sampling for learning to rank. *Inf. Sci.*, 601:90–113. DOI: 10.1016/j.ins.2022.04.012.
- Wang, T., Zhao, X., Lv, Q., Hu, B., and 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.
- Wang, X., Tian, J., Gui, M., Li, Z., Ye, J., Yan, M., and Xiao, Y. (2022). PromptMNER: Prompt-Based Entity-Related Visual Clue Extraction and Integration for Multimodal Named Entity Recognition. In Bhattacharya, A., Lee Mong Li, J., Agrawal, D., Reddy, P. K., Mohania, M., Mondal, A., Goyal, V., and Uday Kiran, R., editors, *Database Systems for Advanced Applications*, pages 297–305, Cham. Springer International Publishing.
- Zhang, S., He, L., Vucetic, S., and Dragut, E. (2018). Regular Expression Guided Entity Mention Mining from Noisy Web Data. In *Empirical Methods in Natural Language Processing*, pages 1991–2000.
- Zhu, Y., Ye, Y., Li, M., Zhang, J., and Wu, O. (2022). Investigating annotation noise for named entity recognition. *Neural Computing and Applications*, 35. DOI: 10.1007/s00521-022-07733-0.