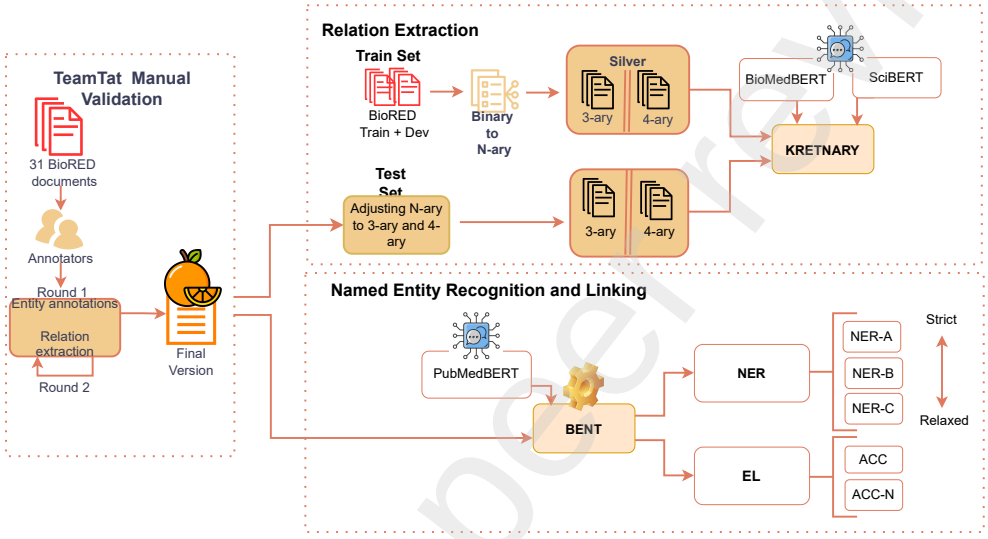
Graphical Abstract

BIORANGE: augmenting BioRED dataset with NIL entity annotation and n-ary relations

Sofia I. R. Conceição, Pedro Ruas, João Fernandes, Francisco M. Couto



Highlights

BIORANGE: augmenting BioRED dataset with NIL entity annotation and n-ary relations

Sofia I. R. Conceição, Pedro Ruas, João Fernandes, Francisco M. Couto

- BIORANGE is a publicly available dataset including NIL entity and n-ary relation annotations.
- Straightforward method for converting existing binary datasets to n-ary.
- Novel silver standard train set for 3 and 4-ary biomedical relations.
- Baseline approaches for NIL entity linking and n-ary relation extraction.

BIORANGE: augmenting BioRED dataset with NIL entity annotation and n-ary relations

Sofia I. R. Conceição[a,*], Pedro Ruas[a], João Fernandes[a], Francisco M. Couto[a]

[a]*LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, , Lisboa, 1749–016, , Portugal*

Abstract

The lack of biomedical datasets focusing on edge cases, such as NIL entities or n-ary relations (i.e. relations involving more than two entities) hinders the development of comprehensive text mining approaches. To address this issue, we have developed guidelines for creating or expanding datasets to include these edge cases. This study demonstrates the potential for reusing existing datasets in a resourceful way to improve edge cases. Our pipeline tackles the scarcity of training data for the extraction of n-ary relations by leveraging 2-ary relations with minimal computational resources. Our results for relation extraction show state-of-the-art performance for 4-ary relations in two BERT-based biomedical models. This approach can be used to augment the value of existing datasets by extracting n-ary relations. As a use case, we provide the gold standard dataset BIORANGE, which results from applying our guidelines to the original BioRED dataset. The expanded dataset includes four additional entity types (*CellTypeOrAnatomicConcept*, *NILGene*, *NILDis*, *NILChem*) totalling 346 new annotations and 79 variable n-ary relation annotations across 31 documents (PubMed titles and abstracts). The dataset is publicly available at: `https://github.com/lasigeBioTM/BIORANGE`.

Keywords: biomedical information extraction, named entity linking, n-ary relation extraction, text mining

*Corresponding author

Email address: `sconceicao@lasige.di.fc.ul.pt` (Sofia I. R. Conceição)

1. Introduction

Scientific literature is the main source for sharing information, particularly in the biomedical domain. Text mining assumes relevance in extracting information stored in the articles' text to translate it into structured, computer-readable knowledge organization systems, such as ontologies, knowledge graphs, and vocabularies. Text mining pipelines usually include tasks such as Named Entity Recognition (NER), Entity Linking (EL), Relation Extraction (RE), Question Answering, among others [1].

Task-specific annotated datasets are useful for training supervised deep-learning-based approaches that have achieved state-of-the-art (SOTA) in several tasks. Human annotation is costly, as it requires time, effort and specific expertise. Additionally, the annotation quality limits the performance of the approaches.

We previously demonstrated the importance of NIL or unlinkable entities to the EL task [2], when new literature is published, the curation process lags behind or Knowledge Organization Systems (KOS) getting outdated, resulting in NIL or unlinkable entities. Several datasets include NIL entities, but these are not associated with any target KOS identifier. They only allow the training and evaluation of approaches that are able to recognize if a given entity is NIL or not. The existence of NIL entities defeats the main purpose of the EL task, since these entities remain locked at text level. In this sense, existing benchmarks hinder the evolution of the EL task into focusing on specific cases. Addressing existing NIL entities has a positive impact on downstream tasks such as Relation Extraction (RE). Finding relations between entities about which we know nothing is significantly more challenging than working with entities that are linked to a standardized resource.

RE is essential to identity associations between entities. It can be applied in real world applications such as drug discovery, by helping discovering protein-protein an drug-target interactions.

Commonly used corpora and relation extraction approaches only consider binary relations (for a list of existing RE biomedical datasets, see BioRED[3]) Table 2. However, additional relevant entities may be necessary to fully characterize a relationship. N-ary relation extraction can help to answer more specific questions such as: given a mutation in a gene, which drug would it respond to, resulting in a gene-mutation-drug, ternary relation [4]; given a gene variation, how does it impact drug response phenotype, a ternary relation of gene variation-drug-phenotype [5]; which type of drug combinations

will result in a positive effect [6] and; given a specific mutation in a gene, how does it affect the reaction to the drug [7].

This work emphasizes the potential for reusing existing datasets in a resourceful manner. By leveraging binary datasets, we aim to maximize the value of existing data. This approach opens up new opportunities for extracting different types of relations, such as n-ary relations. Inspired by recent studies, our hypothesis is that is feasible to transforming lower-arity datasets into n-ary datasets, thereby reutilizing available information to uncover more well-characterized and complex relations.

The fact that many existing datasets are often limited to standard tasks and overlook edge cases hinders the evolution of biomedical text mining. These edge cases, particularly those involving NIL entities and n-ary relations, are typically neglected due to their complexity, impeding the exploration of potentially relevant information.

In order to overcome the scarcity of datasets that include the mentioned edge cases, our main goal is to improve biomedical information extraction pipelines by defining guidelines for annotating NIL entities in addition to extracting n-ary relations. To showcase the usefulness of the guidelines, we augmented the BioRED dataset [3] with annotations for NIL entities and n-ary relations and used machine learning methods to make predictions on both tasks.

The contributions of this work are the following:

- Publicly available dataset including NIL entity and n-ary relation annotations[1]:
  - Four additional entity types compared with the original BioRED dataset: CellTypeOrAnatomicConcept, NILGene, NILDis, NILChem
  - Gold standard with new 341 entity and 79 relations of variable n-ary
  - Silver standard train set for 3 and 4-ary
- Baseline approaches (NIL entity linking and n-ary relation extraction)
  - Leveraging of binary datasets to n-ary train set
- Guidelines to future annotation of additional datasets.

[1] `https://github.com/lasigeBioTM/BIORANGE`

Statement of Significance

Problem	Current biomedical text mining approaches have limitations handling NIL (unlinkable) entities and scarcely considering n-ary (more than two entities) relations.
What is Already Known	Although there have been many prior explorations regarding these edge cases, the existing datasets are often limited to standard tasks and overlook edge cases due to their complexity.
What This Paper Adds	This study proposes leveraging existing datasets by providing guidelines for solving NIL entities and extracting n-ary relations. It demonstrates how to leverage BioRED, a binary dataset, to extract more complex relationships, achieving state-of-the-art performance for 4-ary relations using a simple straightforward method. It also introduces BIORANGE, a publicly available dataset that includes new annotations for NIL entities and n-ary relations.
Who would benefit from the knowledge in this paper	Researchers in the biomedical information extraction field who aim to increase the comprehensiveness and quality of their datasets, as well as the efficacy of their models in extracting complex relationships.

1.1. Related Work

1.1.1. Binary Biomedical Corpora

BioRED[2] [3] is a manually curated dataset of multi-annotated biomedical entities such as Disease or phenotypic feature, chemical entity, gene or gene product, organism taxon, sequence variant and cell line. It was built using 600 PubMed abstracts and it is useful in several biomedical tasks biomedical text mining tasks, namely NER and RE. Regarding the RE task, the relations are binary and consist of the following pairs: gene-disease, chemical-chemical, disease-chemical, gene-chemical, gene-gene, disease-variant, chemical-variant and variant-variant. This corpus was the basis for the LitCoin Natural Language Processing Challenge[3] and the BioCreative VIII challenge [8].

[2] `https://ftp.ncbi.nlm.nih.gov/pub/lu/BioRED/`

[3] https://www.nasa.gov/directorates/stmd/prizes-challenges-crowdsourcing-program/litcoin-challenge/

Commonly used biomedical evaluation datasets focusing on scientific literature with entity annotations besides BioRED include BC5CDR [9], NCBI Disease [10], LINNAEUS [11], CRAFT [12] and MedMentions [13].

BC5CDR [9] was proposed in 2016 in the context of the BioCreative V competition. It is comprised of the titles and abstracts of 1,500 PubMed articles and 4,409 chemical annotations, 5,818 disease annotations, and 3,116 chemical-disease interactions. The entity annotations are associated with identifiers of the Medical Subject Headings (MeSH). 11 diseases and 280 chemical entities have no MeSH identifier, thus are considered to be NIL.

NCBI Disease [10] comprises titles and abstracts of 793 articles from the PubMed collection, which are annotated with 6,892 disease entities linked to MeSH and the Online Mendelian Inheritance in Man (OMIM) catalog. It includes 83 entities that are not associated with any identifier of the target KOS.

LINNAEUS [11] includes 100 full-text documents from the open-access subset of PMC (PMCOA), and the annotations correspond to species names associated with identifiers of the NCBI Taxonomy.

The CRAFT corpus [12] consists of 97 PubMed full-text articles from PMCOA. It contains several biomedical entity types, such as chemical entities, cells, biological processes, cellular and extracellular components, molecular functions, diseases, chemical reactions, organisms, proteins, biomacromolecular entities, and sequence features, anatomical entities, linked to the following KOS, respectively: Chemical Entities of Biological Interest (CHEBI), Cell Ontology (CL), Gene Ontology Biological Process (GO_BP), Gene Ontology Cellular Component (GO_CC), Gene Ontology Molecular Function (GO_MF), MONDO Disease Ontology (MONDO), Molecular Process Ontology (MOP), NCBI Taxonomy (NCBITaxon), Protein Ontology (PR), Sequence Ontology (SO), and Uberon (UBERON).

The MedMentions corpus [13] includes titles and abstracts from 4,392 PubMed articles and 350,000 annotations of biomedical types linked to concepts in the Unified Medical Language System (UMLS).

BioREX, [14] tries to circumvent the issues related to corpus biases, such as small size and domain-specificity, by combining nine datasets into a large one, taking into account the data heterogeneity. This dataset results in five relation pairs: of gene-gene, gene-chemical, gene-disease, chemical-chemical, and chemical-disease. This framework allows the integration of hererogenous data into a single dataset that has been shown to improve RE systems performance. The information merge allowed the pre-trained model to be more

robust and capable of generalizing RE tasks, such as in the drug-drug n-ary dataset.

1.1.2. N-ary Relation extraction

Since most studies focus on binary relations, the majority of available datasets are also focused on binary relations. Although in the recent years advances have been made to create new biomedical n-ary corpora.

One of the first n-ary corpus is a silver standard drug-gene-mutation dataset designed in the context of molecular tumor boards [4]. This dataset was constructed by filtering from an initial set of approximately 1,000,000 full-text articles from PubMed Central and by applying distant supervision. The final dataset resulted in 3,462 ternary relation instances, with only 59 unique relations. Additionally, the dataset is subdivided into binary sub-relations, including 137,469 drug-gene instances and 3,192 drug-mutation instances [4]. Distant supervision was applied to binary pairs using the Gene Drug Knowledge Database [15] and the Clinical Interpretations of Variants In Cancer (CIVIC)[4]. The authors proposed a cross-sentence approach based on a minimal span, where a candidate is retained if there is no other co-occurrence of the same entities in an overlapping text span within fewer consecutive sentences [4].

Another relevant corpus is the PGxCorpus, a manually annotated dataset focused on pharmacogenomics [5]. It includes 945 sentences, 6,761 annotated entities, 2,871 relations, 10 types of entities, and 7 types of relations. While not specifically built for n-ary relations, 92% of its sentences have three relevant target entities that are associated: genomic factor, chemical and phenotype. [5].

[6] released a drug combination dataset with a variable length of n-ary relations. It is a curated dataset that can have from 2 to 15 drug combination mentions. This dataset includes 900 binary relations, 226 ternary relations, and 122 higher-order relations (4-ary and 5-ary), offering data on the effectiveness of drug combinations in therapy [6].

The DUVEL (Detection of Unique Variant Ensembles in Literature) [16] corpus contains relations from oligogenic variant combinations, consisting of gene-variant-gene-variant relationships (4-ary), including cross-sentence annotations, and with binary relation types. It includes 85 PMCO full-text

[4]http://civic.genome.wustl.edu

Table 1: Overview of biomedical corpora.

Dataset	Type	Text Source	NER	NIL	RE
BioRED	Gold	600 PubMed T&ABS	Yes	Yes	2-ary
BC5CDR	Gold	1,500 PubMed T&ABS	Yes	Yes	2-ary
NCBI Disease	Gold	793 PubMed T&ABS	Yes	Yes	No
LINNAEUS	Gold	100 Full-text PMCOA	Yes	Yes	No
CRAFT	Gold	97 PMCOA	Yes	?	No
MedMentions	Gold	4,392 PubMed T&ABS	Yes	Yes	No
BioREX	Silver	BioRED + 8 corpora	Yes	Yes	2-ary
N-ary	Silver	1M Full-text PMCOA	No	No	2- & 3-ary
PGxCorpus	Gold	911 PubMed T&ABS	No	No	3-ary
Drug Combination	Gold	1634 PubMed ABS	No	No	n-ary
DUVEL	Silver	81 PCMO	Yes	No	4-ary
EnzChemRED	Gold	1,210 PubMed ABS	Yes	No	2-ary & 3-ary

T&ABS: Title and Abstract

articles, curated by a single senior annotator.

Recently, EnzChemRED (Enzyme Chemistry Relation Extraction Dataset) dataset [17] identifies 2-ary relations of chemical-chemical and 3-ary relations of protein-(chemical-chemical) that connect binary chemical reactants and the enzymes that catalyze their conversion. This dataset was created from 1,210 PubMed abstracts that were manually curated. It employs single-sentence annotations and supports both binary and multi-class relation types.

Despite these advancements, some datasets still face challenges due to the limited availability of labels, which can hinder supervised learning approaches [18]. To surpass this, [18] developed a method to alleviate the sparsity problem of n-ary relation extraction by decomposing higher-arities in lower-arities and learning the representations to score a new n-ary fact.

Table 1 summarizes the corpora information of this section. N-ary datasets are limited in quantity and size. To overcome this issue, we propose expanding commonly used binary datasets to expand their size and scope, thereby enriching the available data by building upon and enhancing current resources.

2. Methodology

To address our goal of dealing with these edge cases, we expanded a subset of the existing BioRED corpus, including adding a new entity type,

associating NIL entities with target KOS identifiers, and expanding the relation annotations to n-ary relation annotations. We also used the remaining BioRED documents to build a silver standard corpus for the n-ary relation extraction train. An overview of the applied methodology in this work and the differences in the obtained dataset to the original BioRED dataset is presented at 1. In the first phase, TeamTat is used for manual annotation and in the second phase, entities and relations are predicted using machine learning techniques.

2.1. Annotation Guidelines

To the best of our knowledge, no guidelines in the literature are available to annotate NIL entities and n-ary relations. Thus, we created specific guidelines to lead the annotation process specific to this dataset.

2.2. Document selection

The criterion for document selection was the presence of at least one NIL entity, regardless of its type. This criteria resulted in the selection of 31 documents out of the 600 included in the original BioRED.

2.3. Annotators

In all rounds, a manual annotation was performed by 5 annotators with expertise in the fields of biology, biochemistry, bioinformatics and computer science.

2.4. Annotation tasks

The selected documents for annotation retained the original entity types and annotations of the BioRED corpus. Each document underwent two annotation rounds.

The first round consisted of three tasks. The first task was to validate the correctness of the existing annotations and further annotate the present NIL entities with identifiers from target KOS. The second task was to expand the entities, detecting nested or contiguous entities with the new entity type (*CellTypeOrAnatomicConcept)* (CAC). Lastly, the third task consisted of detecting n-ary relations among sentences. The second round aimed to standardize the annotations and resolve any existing ties or disagreements. The specific descriptions of entities and relation annotations are presented in the following subsections.

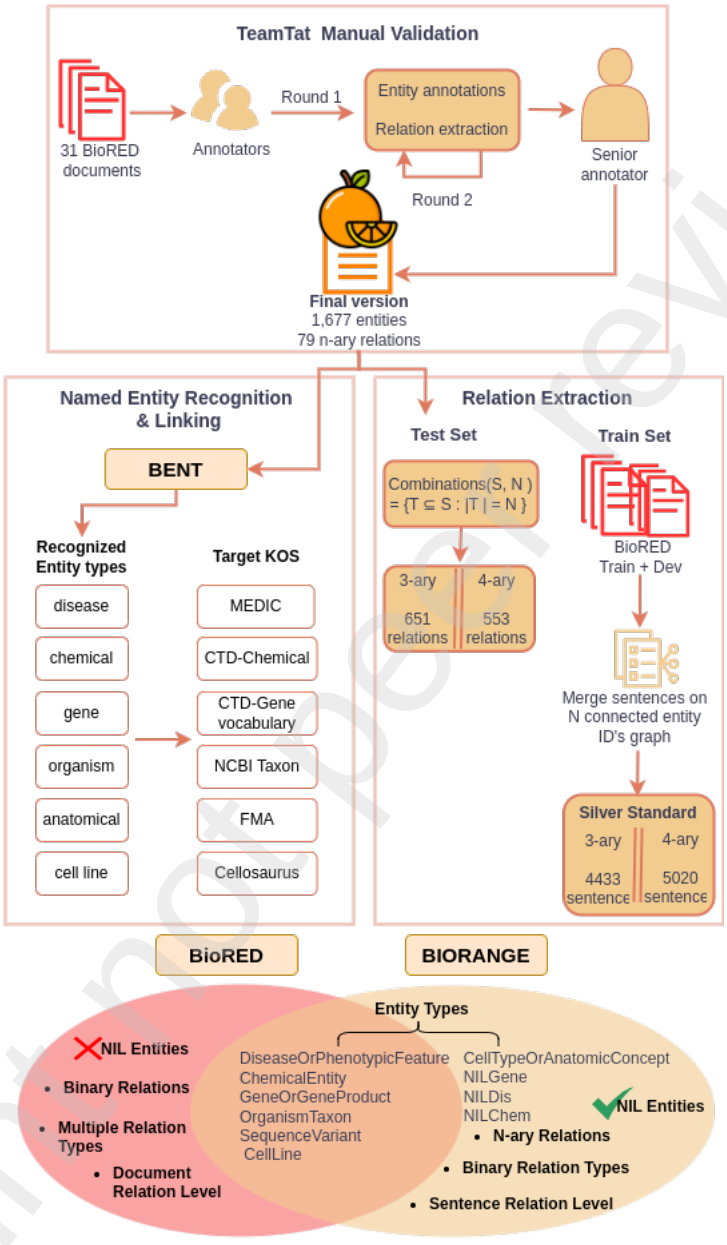


Figure 1: Methodology overview

2.4.1. Entity annotation

The following entity categories present in the BioRed dataset were considered: *DiseaseOrPhenotypicFeature* (DPF), *ChemicalEntity* (CE), *Gene-*

OrGeneProduct (GGP), *OrganismTaxon* (OT), *SequenceVariant* (SV), *CellLine* (CL). The entities belonging to the types DPF, CE and GGP without a valid KB identifier were converted into entities of the types NILDis, NILChem and NILGene, respectively. Besides these categories, the additional category CAC was added to provide additional anatomical context.

The following KOS were used (respective entity type in parenthesis): MEDIC vocabulary [19] (DPF), CTD-Chemical vocabulary [19] (CE), CTD-Gene vocabulary [19] (GGP), NCBI Taxon [20] (OT), dbVar [21] (SV), Cellosaurus [22] (CL), *Foundational Model of Anatomy* ontology, abbreviated by FMA [23] (CAC). The latest versions of the KOS available at the end of August 2024 were used.

2.4.2. Relation annotation

The goal for this task was to expand the original binary relations to n-ary. We follow the assumption that most n-ary relations can be decomposed into k-ary relations that are implied by the n-ary relation [18]. Contrary to the original BioRED that considered the full document for the relations, our approach focuses only on relations between entities in co-occurence in a single sentence. This sentence-level approach provides more precise contextual information [14]. Additionally, work on other datasets, such as the drug-combo dataset [6], shows that in 97% of the abstracts, all drugs involved in a combination attempt are found within a single sentence. Furthermore, for simplification, we considered only binary labels (positive or negative) between any type of entity. The test set for this task was manually validated using only the selected abstracts. Training and validation sets were created using the BioRED version that was made available at the BioCreative VIII [5]. For this seed work, we only considered ternary and quaternary relations.

Training Set: For the creation of the training set, the following steps were done: First, the BioRED train and development sets were combined into one, and then all the 31 PMID's used in our approach were removed.

Using the provided scripts to generate the original BioRED[6], the entities were annotated, and then the abstracts were divided into single sentences. Then, the same sentences with different tagged entities were selected, and using the entity identifiers and the relations, an undirected graph was created.

[5] https://biocreative.bioinformatics.udel.edu/news/biocreative-viii/track-1/

[6] https://ftp.ncbi.nlm.nih.gov/pub/lu/BioRED/

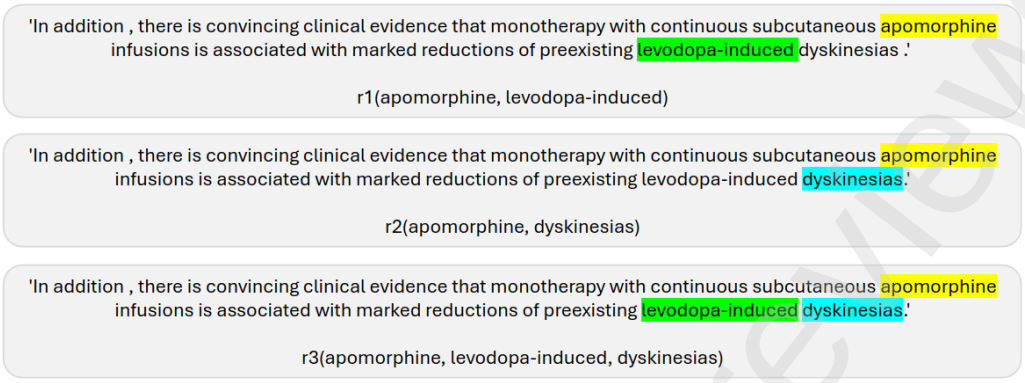


Figure 2: Train set example sentence of binary to ternary relation.

If the graph between the entities was connected, then it was assumed a pathway to the n-ary relation, i.e., a true relationship if at least the N entities (three or four) could be connected. For the negative instances, we randomly sampled triples without interaction. This dataset was divided into 90% for training and 10% for the evaluation set. The resulting train set for tenary relations has 4,433 sentences with 2,315 positive labels and 2,118 negative labels. The quartenary trainset resulted in 5,020 sentences with 4,452 positive labels and 568 negative labels.

Test Set: Only the manual annotations from the senior annotator with experience in RE were used. A initial set of 79 n-ary instances was divided into smaller n-ary (3 and 4-ary). We considered unique triple combinations, given a set of entities in a relation $S = \{e1, e2, e3, .., en\}$, where each subset contains exactly three or four elements. The collection of all subsets for n-ary is expressed as:

$$\text{Combinations}(S, N) = \{T \subseteq S : |T| = N\} \tag{1}$$

Where T is any subset of S with exactly N entities.

Due to the lack of negative labels, similar to the training set, relations with negative labels were generated by choosing triples without interactions, with a proportion of 2:1 (positive:negative) labels. The ternary test set consists of 651 relations: 434 with positive labels and 217 with negative labels. The quaternary test set was created in the same way, taking into account four entities. A test set of 553 relations was generated, including 369 positive and 184 negative relations.

2.4.3. Inter-annotator agreement

Following previous works [24, 25, 5], we determined the inter-annotator agreement using the F1-score. To calculate the pairwise and the overall F1 scores, we compared the annotations of the annotators against the annotations of the corresponding senior annotator.

Hripcsak and Rothschild [24] proposed the following formula for the F1-score:

$$F1 = \frac{2 \times TP}{(2 \times TP + FP + FN)} \tag{2}$$

With TP, FP and FN representing *true positives*, *false positives*, and *false negatives*, respectively. *True positives* occur when a given annotation matches a gold standard annotation, i.e., it is annotated with the same text and span, the same type, and the same KOS identifier. *False positives* occur when a given annotation partially matches the gold standard (i.e. the annotation differs in one of the referred aspects; for example, there is an annotation in the gold standard with the same text and span, same type, but with a different KOS identifier) or does not match the gold standard at all. *False negatives* occur when an annotation present in the gold standard was not annotated.

To determine the agreement with more granularity, we focused on:

a) Span evaluation: Exact match on text, type and span (NER), i.e., strict evaluation of NER and exact match on type and partial match on span and text ($NER*$), i.e., approximate evaluation of NER
b) Identifier evaluation: Exact match on type and KOS identifier (EL), i.e. strict evaluation of EL

The F1-score was first calculated pairwise by entity type, comparing the annotations of each annotator with the reference annotations of the senior annotator. Then, the pairwise F1-scores were averaged.

2.4.4. Annotation tool

To perform the annotation we used the TeamTat web-based tool [26]. The TeamTat tool provided a user-friendly platform to annotate the same documents anonymously to prevent bias, it also provided additional features such as inter-annotator agreement statistics and task completion.

2.5. Automatic entity annotation

BENT[7] is a Python package for entity annotation focused on biomedical text. It performs NER and NEL.

The NER module includes several models based on the pre-trained language model PubMedBERT [27]. PubMedBERT was fine-tuned in specific datasets according to the entity type. The current version includes trained models to recognize the following entity types (the descriptions of the respective training datasets are available in the respective links): "Gene"[8], "Disease"[9], "Bioprocess"[10], "Organism"[11], "Anatomical"[12], "Cell-component"[13], "Cell-type"[14], "Cell-line"[15], "Variant"[16], "Chemical"[17].

The issues of overlaping annotations having different entity types are solved by rule-based module, that includes entity frequency dictionaries determined in the dataset provided by Pubtator Central (file `bioconcepts2-pubtatorcentral` publicly available[18]). The probability of each entity is calculated by determining the frequency of a given entity string in the dataset and the total number of entities present. If the probability calculated for two entities is the same, BENT calculates the frequency of the entities within the document where they are present.

The EL module corresponds to the graph-based model described in [2]. This approach is based on the Personalized PageRank (PPR) algorithm and on the concept of information content. For every input entity in a given document, the candidate generation approach retrieves candidates from the target KOS based on the lexical similarity (Levensthein distance). If the candidate list is empty (i.e., the entity is NIL), the NILINKER model is applied. Then, a disambiguation graph is built for every document, in which the nodes are the candidates for the entities and the edges are added according to the rela-

[7] `https://pypi.org/project/bent/`
[8] `https://huggingface.co/pruas/BENT-PubMedBERT-NER-Gene`
[9] `https://huggingface.co/pruas/BENT-PubMedBERT-NER-Disease`
[10] `https://huggingface.co/pruas/BENT-PubMedBERT-NER-Bioprocess`
[11] `https://huggingface.co/pruas/BENT-PubMedBERT-NER-Organism`
[12] `https://huggingface.co/pruas/BENT-PubMedBERT-NER-Anatomical`
[13] `https://huggingface.co/pruas/BENT-PubMedBERT-NER-Cell-component`
[14] `https://huggingface.co/pruas/BENT-PubMedBERT-NER-Cell-type`
[15] `https://huggingface.co/pruas/BENT-PubMedBERT-NER-Cell-line`
[16] `https://huggingface.co/pruas/BENT-PubMedBERT-NER-Variant`
[17] `https://huggingface.co/pruas/BENT-PubMedBERT-NER-Chemical`
[18] `https://ftp.ncbi.nlm.nih.gov/pub/lu/PubTatorCentral/`

tions between candidates described in the target KOS. The PPR algorithm simulates random walks in the graph to calculate the coherence of each node in the graph: nodes more connected will be more coherent to the graph, thus will be assigned a higher score. The highest-scoring candidate for each input entity is selected.

Evaluation: for NER it was used the tool "brateval" [28]. This tool allows the evaluation of NER performance by calculating four metrics:

- NER-A: strict NER performance, where two annotations (one predicted by BENT, the other present in the gold standard) completely match in terms of span and entity type.
- NER-B: two annotations match if the respective spans overlap and the entity types are shared.
- NER-C: two annotations match if the spans overlap, irrespectively of the entity types.

For EL, we developed an evaluation script to calculate the *accuracy* and the related metric *accuracy-neighboors*. *Accuracy* corresponds to the proportion of correctly annotated entities over the total number of evaluated entities. The only difference to *accuracy-neighboors* is that, in this metric, an annotation is considered to be correctly annotated if the predicted identifier corresponds to the gold standard identifier or, alternatively, to an identifier associated with the direct ancestors or descendants of the entry in the target KOS.

2.6. Automatic relation extraction

The K-RET: Knowledgeable Biomedical Relation Extraction System [29] is a relation extraction system that allows the flexible integration of diverse sources of domain knowledge in the form of ontologies to enhance BERT-based models. However, this system only performs binary relation extraction, focusing on relationships between two entities. To expand its capabilities, the model was adapted to perform n-ary relation extraction, resulting in the KRETNARY system[19].

[19] `https://github.com/lasigeBioTM/KRETNARY`

2.6.1. Hyperparameters and fine-tuning details

All models were trained on eight Tesla M10 GPUs, and the used hyperparameters for KRETNARY runs were a batch size of 16, 20 epochs, class weights of 0.8 and 0.2, and contextual knowledge. The fine-tuned BERT models were `allenai/scibert_base_uncased` [20] and `microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext` [21]. SciBERT [30], is pretrained on scientific text obtained from Semantic Scholar. BiomedBERT [27], previously known as "PubMedBERT", is pretrained using full-text articles from PubMedCentral and abstracts from PubMed. As external knowledge, four sources were used: Gene Ontology (GO) [31, 32], Chemical Entities of Biological Interest (ChEBI) [33], Human Phenotype Ontology (HPO) [34] and Human Disease Ontology (DO) [35].

3. Results and Discussion

The final dataset after TeamTat annotation consisted of 31 annotated documents (titles and respective abstracts) with 1,664 entities and 79 relations. Among the original 79 relations, a total of 33 were 3-ary, 26 4-ary, 13 5-ary, and 7 more than 5-ary.

3.1. Entity annotation

Among the 1,664 entities, 341 correspond to new annotations. Table 2 highlights the distribution of the different types of annotations.

The most frequent entity types are GGP (28.73%), CE (19.89%) and DPF (17.12%), whereas the less frequent type is *NILDis* (0.5%).

The values for the inter-annotator agreement in the entity annotation are shown in Table 3.

The weighted average agreement for both NER, strict and approximate, and EL annotation tasks is, respectively 0.781, 0.786, 0.743. While there is no universal consensus on what constitutes an acceptable agreement level, we reference the thresholds proposed by [25] for a similar evaluation metric (absolute agreement). Based on these thresholds, the observed values for NER exceed the lower limit of acceptable agreement (0.75) and the value for EL gets close.

[20] `https://huggingface.co/allenai/scibert_scivocab_uncased`

[21] `https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract\allowbreak-fulltext`

Table 2: Distribution of annotations by entity type in the BiORANGE dataset, along with the respective percentages relative to to the total number of entities.

Entity type	**Annotations**	**%**
DPF	285	17.12
CE	331	19.89
GGP	478	28.73
OT	167	10.04
SV	50	3.00
CL	12	0.72
CAC	242	14.54
NILGene	25	1.50
NILDis	8	0.5
NILChem	66	3.97
Total	1,664	100.0

3.2. Automatic entity annotation

The performance of the baseline BENT for the NER and EL tasks evaluated in the created dataset is shown in Table 4.

As expected, the performance of BENT in the NER task (Table 4) was the lowest when considering the strictest criteria (NER-A), with an average F1 of 0.705. On the other hand, when considering the strictest criteria (NER-A), BENT obtained the best performance for entities of type OT (0.812) and GGP (0.784) and the worst performance in entities of type CL (0.267) and CAC (0.471).

The F1-scores increased for every entity type when relaxing the evaluation criteria (NER-A to NER-B), which suggests that span mismatches were a major cause of errors performed by BENT. BENT is able to pinpoint the parts of the text that are associated with entities but struggles in determining the exact position where those entities appear. This was particularly evident for entities of type CL, whose performance increased from a F1-score of 0.267 to 0.600, which represents an increase of 0.333.

The increase in performance in NER-C also suggests that another major source of errors is related to the wrong typing of the recognized entities. For instance, the performance when annotating entities of type CL increased 0.141 switching the criteria from NER-B to NER-C. Also, the performance for entities of type CE increases by 0.122, which is related to the fact that often BENT categorizes these entities as GGP. In short, the span of these

Table 3: Overall average agreement (F1-score) for each entity type, assessed using both span-based (NER) evaluation methods (strict and approximate) and identifier-based (EL) evaluation. The averages are computed from pairwise agreements between the three annotators and the reference annotator. Additionally, the weighted average—adjusted for the proportion of each entity type in the dataset—is provided.

	NER		**EL**
Entity Type	Strict	Approx	Strict
CAC	0.211	0.260	0.151
NILDis	0.632	0.632	0.103
NILChem	0.633	0.656	0.123
NILGene	0.577	0.577	0.018
CE	0.949	0.951	0.915
DPF	0.931	0.942	0.896
GGP	0.965	0.976	0.711
SV	0.586	0.605	0.598
CL	0.727	0.727	0.733
OT	0.938	0.938	0.929
Weighted average	0.781	0.786	0.743

entities is accurately or approximately recognized, but they are classified with the wrong entity type.

The average performance of BENT in the EL task (Table 4) was 0.573 when considering the strictest criteria and 0.669 when additionally considering neighbors in the evaluation. The performance was higher for entities of type CE (0.825) and DPF (0.698) and lower for entities of type NILChemical (0.197) and NILGene (0.200). Dealing with NIL entities represents a major challenge since it is not clear to which entry in the target KOS a given approach should map these entities. With the present work, we attempted to overcome a limitation identified in the existing literature through the definition of guidelines to annotate these types of entities. However, the low agreement obtained between the different annotators (Table 3) suggests that there is still room for improvement.

The increase in the performance achieved by BENT when relaxing the evaluation criteria (Accuracy-neighbours) demonstrates that a major source of errors is related to the specificity of the annotations. The annotation guidelines greatly vary between evaluation datasets, which originates different interpretations about the specificity of entities. In some cases, BENT

Table 4: BENT performance in the NER and EL tasks.

	GGP	CE	DPF	CAC	OT	CL	NILChem	NILGene
NER-A								
P (Avg 0.745)	0.778	0.679	0.831	0.571	0.904	0.222	-	-
R (Avg 0.668)	0.791	0.613	0.726	0.401	0.737	0.333	-	-
F1 (Avg 0.705)	0.784	0.644	0.775	0.471	0.812	0.267	-	-
NER-B								
P (Avg 0.828)	0.846	0.753	0.920	0.688	0.978	0.500	-	-
R (Avg 0.742)	0.860	0.680	0.804	0.484	0.796	0.750	-	-
F1 (Avg 0.783)	0.853	0.714	0.858	0.568	0.878	0.600	-	-
NER-C								
P (Avg 0.903)	0.914	0.911	0.959	0.744	0.978	0.667	-	-
R (Avg 0.776)	0.863	0.773	0.821	0.541	0.790	0.833	-	-
F1 (Avg 0.834)	0.887	0.837	0.885	0.627	0.874	0.741	-	-
EL Task								
Accuracy (Avg 0.573)	0.479	0.825	0.698	0.430	0.539	-	0.197	0.200
Accuracy-neighboors (Avg 0.669)	-	0.833	0.765	0.484	0.593	-	0.318	-

links a recognized entity to either a direct descendant or a parent in the target KOS instead of the correct entry. BENT is not a supervised approach trained on a specific dataset, so we predict that this type of error will persist. Two examples illustrate this type of error: in the document 24442316, BENT linked the entity *"liver injured"* to the entry *"Chemical and Drug Induced Liver Injury"* (identifier MESH:D056486), but the correct entry would be *"Liver Diseases"* (identifier MESH:D008107), which is the parent of *"Chemical and Drug Induced Liver Injury"*; in the document 17879945, BENT linked the entity *"female hormones"* to the entry *"Gonadal Hormones"* (identifier MESH:D042341) instead of the entry *"Hormones"* (identifier MESH:D006728), which is the parent of *"Gonadal Hormones"*.

3.3. RE Inter-annotator agreement

All relation IDs were normalized and compared. Two different comparisons were made: full match, where there is a full match between all entities present on the relation, and partial match, where sub-relations of the main relation were considered. The annotations made by the most senior annotator were considered the gold standard for comparison.

Table 5 presents the scores for both exact and partial matches among all documents, being the first row of each annotator regarding the 1 round and the second regarding the 2 round. The partial match scores improved from the exact match scores, despite the relatively low scores, being the highest F1 achieved by annotator 1 on R1 0.518. Overall, the RE inter-annotator

Table 5: RE Inter-annotator Agreement: exact and partial match scores.

	Exact Match			Partial Match		
Annotator	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Annotator 1	0.261	0.536	0.351	0.424	0.652	0.514
	0.271	0.257	0.264	0.429	0.354	0.388
Annotator 2	0.230	0.750	0.352	0.307	0.800	0.444
	0.217	0.407	0.283	0.295	0.484	0.367
Annotator 3	0.138	0.434	0.210	0.147	0.448	0.221
	0.136	0.196	0.160	0.141	0.201	0.166
Annotator 4	0.132	0.156	0.143	0.197	0.217	0.207
	0.094	0.117	0.104	0.169	0.193	0.180

agreement scores show that there was a low level of consistency on this task among the different annotators.

The scores regarding RE are dependent on the NER phase, since each annotator performed their own NER. If the annotators selected different entities (for example, entity boundary disagreements or nested entities), it would have a negative impact on the agreement. Furthermore, RE depends on the linguist interpretation, subjectivity, and expertise of each annotator, which increases the level of disagreement. The complexity of the relations' variable arity also creates an additional issue in that the annotators may not always select all the available entity options for the relation or select a different criteria to define a relation.

3.4. Relation Prediction

The results for the 3-ary and 4-ary datasets are presented in Table 6. The results show that SciBERT has a consistent better performance in comparison to BiomedBERT. This model seems to perform better on the 4-ary dataset, where there is an increase of 0.184 in the macro average F1 score from 3-ary to 4-ary. Both models struggle to predict false labels on both test sets, but SciBERT is better at identifying true positives.

In the BioREx paper [14] it was performed an evaluation on the drug-drug N-ary combination dataset [6], achieving SOTA results with the exact match F1-score of 66.2% for the positive combination and 75.8% for any combination. Additionally, the higher F1-score on the DUVEL dataset [16] was 84% and for 3-ary results, a F1-score of 83.8% on binary class is achieved in the EnzChemRED work [17]. We achieve SOTA results for the 4-ary. Although

Table 6: Performance metrics for all finetuned models on n-ary test sets.

	SciBERT			BiomedBERT			Binary SciBERT		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
3-ary	Accuracy 0.677			Accuracy 0.616			Accuracy 0.608		
Label False	0.515	0.544	0.529	0.449	0.673	0.539	0.135	0.032	0.052
Label True	0.765	0.744	0.755	0.782	0.588	0.671	0.649	0.896	0.753
Macro Avg	0.640	0.644	0.642	0.615	0.630	0.605	0.459	0.464	0.403
4-ary	Accuracy 0.854			Accuracy 0.798			Accuracy 0.638		
Label False	0.841	0.690	0.758	0.671	0.766	0.716	0.278	0.054	0.091
Label True	0.858	0.935	0.895	0.875	0.813	0.843	0.663	0.930	0.774
Macro Avg	0.849	0.812	0.826	0.773	0.789	0.779	0.471	0.492	0.433

our results do not achieve SOTA in the 3-ary data, they are not negligible, especially considering that we fine-tuned our model using a significantly smaller and less diverse training set.

To ensure consistency with the n-ary SOTA performance achieved using the binary BioREX dataset, we evaluated the 3-ary and 4-ary test sets using a model finetuned on BioRED binary single-sentence data. We finetuned the SciBERT model using a training set of 7,894 sentences and an evaluation set of 879 sentences, applying the same hyperparameters described in 2.6.1. The results shown in Table 6, column Binary SciBERT, indicate that training in binary version alone is not enough to obtain a strong performance on n-ary datasets.

These results demonstrate that combining the existing relations of a gold standard binary dataset and merging them to create an n-ary dataset using internal labels as distant supervision is an effective and efficient method for generating quality n-ary relations to train models.

3.5. Challenges and Bias in N-ary Relation Extraction

3.5.1. N-ary Complexity

When extracting biological relationships, it is essential to consider how the different entities interact within a given context. Considering the following example sentence:

> *"Maleate-induced renal injury included **increase** in renal vascular resistance and in the urinary excretion of total protein, **glucose**, sodium, **neutrophil gelatinase-associated lipocalin***

*(NGAL) and **N-acetyl b-D-glucosaminidase** (NAG), upregulation of kidney injury molecule (KIM)-1, **decrease** in **renal blood flow** and **claudin-2** expression besides of necrosis and apoptosis of tubular cells on 24 h."*

In the example sentence, entities such as *"glucose"*, *"neutrophil gelatinase-associated lipocalin"* and *"N-acetyl b-D-glucosaminidase"* all show an "increased" relationship with "maleate-induced renal injury". Thus, when extracting relationships, these entities can be grouped under the category "increased in maleate-induced renal injury".

However, when considering other entities like *"claudin-2"* and *"renal blood flow"*, these show a "decreased" relationship with maleate-induced renal injury, suggesting that these entities should be grouped separately under the category "decreased in maleate-induced renal injury".

For example, in a 3-ary scenario where the tagged entities in this sentence are *"glucose"*, *"neutrophil gelatinase-associated lipocalin"* and *"claudin-2"*, the relationships become more ambiguous. A relationship's "validity" depends on how we define it. Whether we consider a change in expression or concentration that can be grouped under a single relation or if we simply consider all entities as "altered in maleate-induced renal injury" regardless of the differences of interactions.

This adds a layer of complexity to n-ary relations because the validity of a relationship is determined by the criteria used to define a "valid" relation. This criterion can be subjective and context-dependent, making the extraction process more complicated and more open to interpretation.

Arity Influence Our results demonstrate better performance in predicting 4-ary relations compared to 3-ary relations. One possible explanation is that including a fourth entity adds context, which helps clarify relationships and reduces ambiguity. In sentences with a large number of other non-tagged entities, such as lists of disorders, the presence of a fourth entity frequently clarifies the roles of the entities in the relation. This additional clarity contributes to accurately identifying valid relationships, particularly in complex contexts.

Moreover, fixed arities still pose limitations, as they restrain obtaining the full relation, such as in this example sentence:

*"The most commonly reported toxic effects of **capecitabine** are diarrhea, nausea, **vomiting**, **stomatitis**, and **hand-foot***

syndrome"

To fully characterize the toxic effects of *"capecitabine"*, *"diarrhea"* and *"nausea"* would be required. Since the arity was fixed at 4, only a partial characterization is made. Although this approach captures the relation more effectively than binary methods, it still does not capture the full complexity of the information.

3.5.2. Manual Validation Bias

Reaching a high inter-annotator agreement would be the best approach to mitigate bias and enhance the validity and robustness of the annotations. The low scores and problems discussed in the 3.3 section led to the decision to rely on annotations from a single senior annotator. It is important to acknowledge that this choice introduces potential biases into the dataset. Specifically, the annotations may be influenced by the annotator's individual interpretation of the task, as well as other factors such as annotation fatigue. These factors can result in inconsistencies or subjective judgments that may not be representative of broader perspectives. Regardless of these limitations, the decision to proceed with a single annotator was made to maintain consistency and reduce the complexity introduced by multiple annotators with differing interpretations.

3.5.3. Processing Issues and Error Analysis

Challenges and problems can arise from building n-ary relations in an automatic manner. Extending binary to n-ary relations may introduce false positives [18]. Similar issues can arise when selecting negative samples from entities that are assumed not to interact, resulting in false negatives. Furthermore, negative relations may not represent realistic relations as they fail to represent boundary cases [36].

Errors in sentence splitting, tag addition, and removal can also occur by the automatic separation of n-ary to k-ary relations.

SciBERT Error Analysis Five random samples of false positive (FP) and false negative (FN) of each 3 and 4-ary SciBERT predictions were analyzed. FP predictions for 3-ary and 4-ary fall in the category of ambiguous sentences (example 1). Furthermore, for 3-ary it also occurs in sentences that were incorrectly labeled as false but in fact have a relationship (example 2).

- Example 1: *"Low activity of patient plasma butyrylcholinesterase with* `<e1>` *butyrylthiocholine* `</e1>` *(BTC) and* `<e2>` *benzoylcholine* `</e2>`*,*

and values of dibucaine and <e3> fluoride </e3> numbers fit with heterozygous atypical silent genotype."
"Eight Alu sequences (ACE, <e1> TPA25 </e1>, PV92, <e2> APO </e2>, FXIIIB, <e3> D1 </e3>, <e4> A25 </e4> and B65) were analyzed in two samples from Navarre and Guipuzcoa provinces (Basque Country, Spain)."

- Example 2: *"<e1> Lipoprotein glomerulopathy </e1> (LPG) is a rare disease characterized by the presence of <e2> thrombuslike deposition </e2> in markedly dilated <e3> glomerular capillaries </e3> and is often accompanied by an increased serum apolipoprotein E (apoE) level."*

FN predictions for 3-ary and 4-ary appear to be primarily difficulties with acronyms (example 3).

- Example 3: *"Myotonic dystrophy (<e1> DM </e1>), the most prevalent <e2> muscular disorder </e2> in adults, is caused by (CTG) n-repeat expansion in a gene encoding a protein kinase (DM protein kinase; <e3> DMPK </e3>) and involves changes in cytoarchitecture and ion homeostasis."*
"Of these, <e1> NOX1 </e1> and <e2> NOX2 </e2> have been reported to contribute to intravitreal neovascularization (<e3> IVNV </e3>) in oxygen-induced <e4> retinopathy </e4> (OIR) models.

4. Conclusion

This work highlights the potential for reusing existing datasets, specifically BioRED, in a resourceful manner to improve text mining approaches dealing with edge cases. We present an open-access corpus that has been enhanced with four additional entity types, n-ary relation annotations, and NIL entities. Using a carefully validated dataset, we show that binary datasets can be used to train n-ary models by presenting baseline methods that resulted in a 3-ary and 4-ary silver corpus. Additionally, instructions for annotating other datasets are provided, laying the basis for further study and development.

Our RE results show SOTA values for 4-ary in both SciBERT and BiomedBERT models. Furthermore, the leveraging of binary datasets may be a future direction for maximizing the value of existing datasets by extracting

new types of relations, such as n-ary relations, using minimal computational methods.

Developing guidelines for annotating the edge cases is difficult, as evidenced by the relatively low inter-annotator agreement in both NER and RE tasks. We believe the guidelines will be useful for future annotation tasks.

Future work on the RE task might benefit investigating more rich dataets, as in the case of BioREX, integrating and evaluating NIL entities, and exploring RE without fixed n-arities and with dependency parsing for more accurate relations. In addition, future iterations of this corpus could benefit from a more diverse annotation process involving more senior annotators to improve generalizability.

Declaration of Generative AI and AI-assisted technologies in the writing process

ChatGPT-4 and QuillBot tools were used exclusively for grammar checking and enhancing sentence clarity. All the content and ideas presented in this paper were developed by the authors.

CRediT authorship contribution statement

Sofia I. R. Conceição: Conceptualization, Methodology, Software, Writing – original draft, Data curation, Investigation. **Pedro Ruas:** Conceptualization, methodology, software, Writing – original draft, Data curation, Investigation. **João Fernandes:** Methodology, Writing – review, Data curation. **Francisco M. Couto:** Conceptualization, Writing – review & editing, Funding acquisition, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge the help of Furkan Goz and Filipe Nascimento in the curating phase of both NER, EL and RE tasks. We are grateful to the authors

of BioRED for making the data and codes publicly available. This work was supported by FCT (Fundacão para a Ciência e a Tecnologia) through funding of the PhD Scholarships with ref. 2020.05393.BD attributed to PR and ref. UI/BD/153730/2022 attributed to SIRC, and the LASIGE Research Unit, ref. UID/000408/2025.

References

[1] W. W. Fleuren, W. Alkema, Application of text mining in the biomedical domain, Methods 74 (2015) 97–106, text mining of biomedical literature. doi:https://doi.org/10.1016/j.ymeth.2015.01.015.
URL https://www.sciencedirect.com/science/article/pii/S1046202315000274

[2] P. Ruas, F. M. Couto, Nilinker: Attention-based approach to nil entity linking, Journal of Biomedical Informatics 132 (8 2022). doi:10.1016/j.jbi.2022.104137.

[3] L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, Z. Lu, BioRED: a rich biomedical relation extraction dataset, Briefings in Bioinformatics 23 (5), bbac282 (07 2022). arXiv:https://academic.oup.com/bib/article-pdf/23/5/bbac282/45936115/bbac282.pdf, doi:10.1093/bib/bbac282.
URL https://doi.org/10.1093/bib/bbac282

[4] N. Peng, H. Poon, C. Quirk, K. Toutanova, W. tau Yih, Cross-sentence n-ary relation extraction with graph LSTMs, Transactions of the Association for Computational Linguistics 5 (2017) 101–115. doi:10.1162/tacl_a_00049.

[5] J. Legrand, R. Gogdemir, C. Bousquet, K. Dalleau, M.-D. Devignes, W. Digan, C.-J. Lee, N.-C. Ndiaye, N. Petitpain, P. Ringot, PGxCorpus, a manually annotated corpus for pharmacogenomics, Scientific Data 7 (3) (2020) 1–13. doi:10.1038/s41597-019-0342-9.

[6] A. Tiktinsky, V. Viswanathan, D. Niezni, D. Meron Azagury, Y. Shamay, H. Taub-Tabib, T. Hope, Y. Goldberg, A dataset for n-ary relation extraction of drug combinations, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for

Computational Linguistics, Seattle, United States, 2022, pp. 3190–3203. doi:10.18653/v1/2022.naacl-main.233.

[7] R. Jia, C. Wong, H. Poon, Document-level n-ary relation extraction with multiscale representation learning, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3693–3704. doi:10.18653/v1/N19-1370.

[8] R. Islamaj, Proceedings of the biocreative viii challenge and workshop: Curation and evaluation in the era of generative models, in: Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models, Zenodo, 2023. doi:10.5281/zenodo.10103191.

[9] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, Z. Lu, Biocreative v cdr task corpus: a resource for chemical disease relation extraction, Database 2016 (2016).

[10] R. I. Doğan, R. Leaman, Z. Lu, Ncbi disease corpus: a resource for disease name recognition and concept normalization, Journal of biomedical informatics 47 (2014) 1–10.

[11] M. Gerner, G. Nenadic, C. M. Bergman, Linnaeus: A species name identification system for biomedical literature, BMC Bioinformatics 11 (1) (2010) 85, published on 2010/02/11. doi:10.1186/1471-2105-11-85.
URL https://doi.org/10.1186/1471-2105-11-85

[12] K. B. Cohen, K. Verspoor, K. Fort, C. Funk, M. Bada, M. Palmer, L. E. Hunter, The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation in the Biomedical Domain, Springer Netherlands, Dordrecht, 2017, pp. 1379–1394, https://doi.org/10.1007/978-94-024-0881-2_53. doi:10.1007/978-94-024-0881-2$_5$3.

[13] S. Mohan, D. Li, Medmentions: A large biomedical corpus annotated with UMLS concepts, CoRR abs/1902.09476 (2019). arXiv:1902.09476.
URL http://arxiv.org/abs/1902.09476

[14] P.-T. Lai, C.-H. Wei, L. Luo, Q. Chen, Z. Lu, Biorex: Improving biomedical relation extraction by leveraging heterogeneous datasets, Journal of Biomedical Informatics 146 (2023) 104487. doi:https://doi.org/10.1016/j.jbi.2023.104487.
URL https://www.sciencedirect.com/science/article/pii/S1532046423002083

[15] R. Dienstmann, I. S. Jang, B. Bot, S. Friend, J. Guinney, Database of Genomic Biomarkers for Cancer Drugs and Clinical Targetability in Solid Tumors, Cancer Discovery 5 (2) (2015) 118–123. arXiv:https://aacrjournals.org/cancerdiscovery/article-pdf/5/2/118/1714782/118.pdf, doi:10.1158/2159-8290.CD-14-1118.
URL https://doi.org/10.1158/2159-8290.CD-14-1118

[16] C. Nachtegael, J. De Stefani, A. Cnudde, T. Lenaerts, Duvel: an active-learning annotated biomedical corpus for the recognition of oligogenic combinations, Database 2024 (2024) baae039. arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/baae039/58365505/baae039.pdf, doi:10.1093/database/baae039.
URL https://doi.org/10.1093/database/baae039

[17] P.-T. Lai, E. Coudert, L. Aimo, K. Axelsen, L. Breuza, E. De Castro, M. Feuermann, A. Morgat, L. Pourcel, I. Pedruzzi, et al., Enzchemred, a rich enzyme chemistry relation extraction dataset, Scientific Data 11 (1) (2024) 982.

[18] K. Akimoto, T. Hiraoka, K. Sadamasa, M. Niepert, Cross-sentence n-ary relation extraction using lower-arity universal schemas, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6225–6231. doi:10.18653/v1/D19-1645.
URL https://aclanthology.org/D19-1645

[19] A. P. Davis, T. C. Wiegers, R. J. Johnson, D. Sciaky, J. Wiegers, C. J. Mattingly, Comparative Toxicogenomics Database (CTD): update 2023, Nucleic Acids Research 51 (D1)

(2022) D1257–D1262. arXiv:https://academic.oup.com/nar/article-pdf/51/D1/D1257/48441054/gkac833.pdf, doi:10.1093/nar/gkac833.
URL https://doi.org/10.1093/nar/gkac833

[20] C. L. Schoch, S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, S. Sharma, V. Soussov, J. P. Sullivan, L. Sun, S. Turner, I. Karsch-Mizrachi, NCBI Taxonomy: a comprehensive update on curation, resources and tools, Database 2020 (2020) baaa062. arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/baaa062/33570620/baaa062.pdf, doi:10.1093/database/baaa062.
URL https://doi.org/10.1093/database/baaa062

[21] I. Lappalainen, J. Lopez, L. Skipper, T. Hefferon, J. D. Spalding, J. Garner, C. Chen, M. Maguire, M. Corbett, G. Zhou, J. Paschall, V. Ananiev, P. Flicek, D. M. Church, dbVar and DGVa: public archives for genomic structural variation, Nucleic Acids Research 41 (D1) (2012) D936–D941. arXiv:https://academic.oup.com/nar/article-pdf/41/D1/D936/3686260/gks1213.pdf, doi:10.1093/nar/gks1213.
URL https://doi.org/10.1093/nar/gks1213

[22] A. Bairoch, The Cellosaurus, a Cell-Line Knowledge Resource, Journal of Biomolecular Techniques 29 (2) (2018) 25–38, epub 2018 May 10. doi:10.7171/jbt.18-2902-002.

[23] C. Rosse, J. L. V. Mejino, The Foundational Model of Anatomy Ontology, Springer London, London, 2008, pp. 59–117, https://doi.org/10.1007/978-1-84628-885-2_4. doi:10.1007/978-1-84628-885-2$_4$.

[24] G. Hripcsak, A. S. Rothschild, Agreement, the f-measure, and reliability in information retrieval, Journal of the American Medical Informatics Association 12 (2005) 296–298. doi:10.1197/jamia.M1733.

[25] C. Grouin, S. Rosset, P. Zweigenbaum, K. Fort, O. Galibert, L. Quintard, Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview, in: N. Ide, A. Meyers, S. Pradhan, K. Tomanek (Eds.), Proceedings of the 5th Linguistic Annotation Workshop, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 92–100.
URL https://aclanthology.org/W11-0411/

[26] R. Islamaj, D. Kwon, S. Kim, Z. Lu, TeamTat: a collaborative text annotation tool, Nucleic Acids Research 48 (W1) (2020) W5–W11. arXiv:https://academic.oup.com/nar/article-pdf/48/W1/W5/33433452/gkaa333.pdf, doi:10.1093/nar/gkaa333.
URL https://doi.org/10.1093/nar/gkaa333

[27] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing (2020). arXiv:arXiv:2007.15779.

[28] K. Verspoor, A. Jimeno Yepes, L. Cavedon, T. McIntosh, A. Herten-Crabb, Z. Thomas, J.-P. Plazzer, Annotating the biomedical literature for the human variome, Database 2013 (2013) bat019. arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/bat019/16732660/bat019.pdf, doi:10.1093/database/bat019.
URL https://doi.org/10.1093/database/bat019

[29] D. F. Sousa, F. M. Couto, K-ret: knowledgeable biomedical relation extraction system, Bioinformatics 39 (4) (2023) btad174.

[30] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, in: EMNLP, Association for Computational Linguistics, 2019.
URL https://www.aclweb.org/anthology/D19-1371

[31] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al., Gene ontology: tool for the unification of biology, Nature genetics 25 (1) (2000) 25–29.

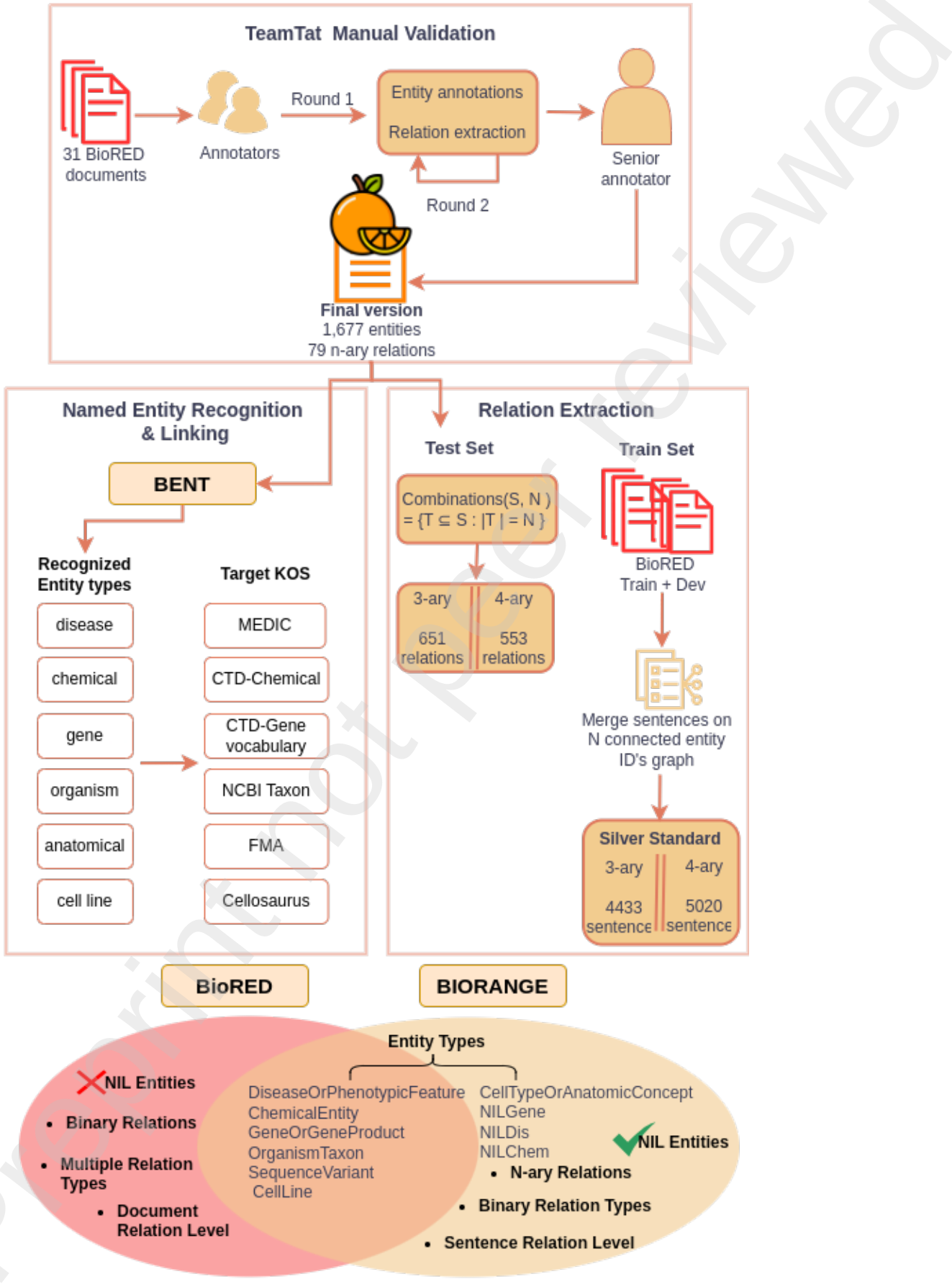
[32] G. O. Consortium, The gene ontology resource: 20 years and still going strong, Nucleic acids research 47 (D1) (2019) D330–D338.

[33] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner, Chebi: a database and ontology for chemical entities of biological interest, Nucleic acids research 36 (suppl_1) (2007) D344–D350.

[34] S. Köhler, M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, D. Danis, G. Balagura, G. Baynam, A. M. Brower, et al., The human phenotype ontology in 2021, Nucleic acids research 49 (D1) (2021) D1207–D1217.

[35] L. M. Schriml, J. B. Munro, M. Schor, D. Olley, C. McCracken, V. Felix, J. A. Baron, R. Jackson, S. M. Bello, C. Bearer, et al., The human disease ontology 2022 update, Nucleic acids research 50 (D1) (2022) D1255–D1261.

[36] Y. Lin, K. Lu, S. Yu, T. Cai, M. Zitnik, Multimodal learning on graphs for disease relation extraction, Journal of Biomedical Informatics 143 (2023) 104415. doi:https://doi.org/10.1016/j.jbi.2023.104415.
URL https://www.sciencedirect.com/science/article/pii/S1532046423001363



'In addition , there is convincing clinical evidence that monotherapy with continuous subcutaneous apomorphine infusions is associated with marked reductions of preexisting levodopa-induced dyskinesias .'

r1(apomorphine, levodopa-induced)

'In addition , there is convincing clinical evidence that monotherapy with continuous subcutaneous apomorphine infusions is associated with marked reductions of preexisting levodopa-induced dyskinesias.'

r2(apomorphine, dyskinesias)

'In addition , there is convincing clinical evidence that monotherapy with continuous subcutaneous apomorphine infusions is associated with marked reductions of preexisting levodopa-induced dyskinesias.'

r3(apomorphine, levodopa-induced, dyskinesias)