Optimizing Relation Extraction in Medical Texts through Active Learning: A Comparative Analysis of Trade-offs

Siting Liang[1], **Pablo Valdunciel Sánchez**[*], **Daniel Sonntag**[1,2]
[1]German Research Center for Artificial Intelligence, Germany
[2]University of Oldenburg, Germany
siting.liang|daniel.sonntag@dfki.de

Abstract

This work explores the effectiveness of employing Clinical BERT for Relation Extraction (RE) tasks in medical texts within an Active Learning (AL) framework. Our main objective is to optimize RE in medical texts through AL while examining the trade-offs between performance and computation time, comparing it with alternative methods like Random Forest and BiLSTM networks. Comparisons extend to feature engineering requirements, performance metrics, and considerations of annotation costs, including AL step times and annotation rates. The utilization of AL strategies aligns with our broader goal of enhancing the efficiency of relation classification models, particularly when dealing with the challenges of annotating complex medical texts in a Human-in-the-Loop (HITL) setting. The results indicate that uncertainty-based sampling achieves comparable performance with significantly fewer annotated samples across three categories of supervised learning methods, thereby reducing annotation costs for clinical and biomedical corpora. While Clinical BERT exhibits clear performance advantages across two different corpora, the trade-off involves longer computation times in interactive annotation processes. In real-world applications, where practical feasibility and timely results are crucial, optimizing this trade-off becomes imperative.

1 Introduction

The digitisation of diverse medical documents into Electronic Health Records (EHRs) has significantly increased worldwide. Essential relationships among biomedical entities, including drug-drug interactions and treatment efficacy lie within EHRs (Herrero-Zazo et al., 2013; Uzuner et al., 2011; Henry et al., 2020). Biomedical and clinical texts often contain complex and highly specialized language, making it difficult for models to understand and extract relationships accurately (Zhou et al., 2014; Bose et al., 2021a). Figure 1 shows the annotated relations between different pairs of named entities. Relation Extraction (RE) systems aim to identify the relevant entity mentions and recognize their relations. Previous research consistently underscores the superior performance of deep learning methods in biomedical and clinical RE tasks within passive learning environments (Wei et al., 2020; Yadav et al., 2022). However, the annotation process required to construct training datasets is both time-consuming and expensive.

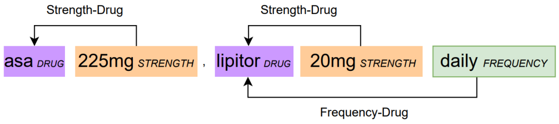


Figure 1: Demonstration of entities and their annotated relations in the n2c2 corpus (Henry et al., 2020): Each instance may feature multiple entities, and the annotations indicate the presence or absence of a relation between any two entities.

Active Learning (AL) advocates for a gradual labelling approach focused on the most informative instances by strategically selecting challenging or uncertain instances (Settles, 2009, 2012; Zhang et al., 2012; Shelmanov et al., 2019). However, utilizing AL with pre-trained language models can result in unacceptable waiting time for annotators (Maekawa et al., 2022). Traditional machine learning (ML) models emerge as potentially more suitable choices for AL settings due to their shorter iteration times and commendable performance on the same tasks (Munkhdalai et al., 2018). This work aims to assess potential variations in annotation costs for biomedical and clinical RE using various supervised learning methods with AL: Random Forest (Alimova and Tutubalina, 2020), Bidirectional Long Short Term Memory (BiLSTM) networks (Hasan et al., 2020), and a pre-trained Clin-

[*] Work completed during master thesis at German Research Center for Artificial Intelligence.

Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024), pages 23–34
March 22, 2024 ©2024 Association for Computational Linguistics

ical BERT model (Alsentzer et al., 2019). The evaluation is conducted on two relevant RE benchmarks, namely the Drug-Drug Interaction (DDI) corpus (Herrero-Zazo et al., 2013) and the n2c2 corpus (Henry et al., 2020). The primary objective is to understand the data requirements for RE in medical text and identify resource-efficient strategies for real-world applications. This investigation aligns with the principles of Interactive Machine Learning (IML) (Amershi et al., 2014; Dudley and Kristensson, 2018; Wang et al., 2021; Wu et al., 2022; Liang et al., 2023). The goal is to enhance model adaptability in the medical domain while minimizing the annotation workload for domain experts, particularly when utilizing pre-trained language models.

The experimental results reveal that three machine learning approaches achieve performance comparable to state-of-the-art methods with significantly less annotated data through active learning (AL), resulting in a substantial reduction in annotation costs. Our analysis, comparing Clinical BERT with alternatives such as Random Forest and BiLSTM networks, offers insights into the advantages and challenges of employing AL strategies with advanced pre-trained language models. This study contributes to optimizing relation extraction in medical texts by exploring the trade-offs of different ML methods within an AL framework.

2 Approach

2.1 Data and Classification Scheme

DDI Corpus. The DDI corpus (Herrero-Zazo et al., 2013) comprises 792 documents describing short drug-drug interactions (DDIs) from the DrugBank database (DDI-DrugBank corpus) and 233 MedLine abstracts (DDI-MedLine corpus), with four clinical entity types, namely *Drug*, *Brand*, *Group* and *Drug_n*. The corpus proposes four types of DDI relations: *Effect*, *Mechanism*, *Advise* and *Interaction*. Table 9 reports the frequency counts of the different relation types in the corpus, where *Effect* denotes the description of the effect of the drug-drug interaction, *Mechanism* is assigned when a pharmacodynamic or pharmacokinetic interaction occurs, *Advise* is assigned to those drug-drug interactions that provide recommendations or advice regarding their concomitant use and *Interaction* is assigned when the sentence merely states that interaction occurs without providing additional information about the interaction. An example of each relation type is illustrated in Table 1. The corpus can be accessed from the official GitHub repository [1]. For DDI corpus, a multi-class classification scheme is proposed in previous work, where one classifier determines one possible drug-drug interaction or no relation between two target entities. Examples of each relation type are presented in Table 1. More data statistics on the sizes of the datasets and average sequence lengths can be found in the appendix A.

Relation Type	Example
Advise	Interactions may be expected, and [UROXATRAL]$_{BRAND}$ should NOT be used in combination with other [alpha-blockers]$_{GROUP}$.
Effect	In common with other broad-spectrum antibiotics, [AUGMENTIN XR]$_{BRAND}$ may reduce the efficacy of oral [contraceptives]$_{GROUP}$.
Mechanism	Milk, milk products, and [calcium]$_{DRUG}$ -rich foods or drugs may impair the absorption of [EMCYT]$_{BRAND}$.
Int	Conversely, [diethylpropion]$_{DRUG}$ may interfere with [antihypertensive drugs]$_{GROUP}$.

Table 1: Instances of relation types annotated to pairs of entities in the DDI corpus.

n2c2 Corpus. The n2c2 corpus is specifically designed for the medication challenge and emphasises the identification of injuries caused by drug-related medical interventions, such as allergic reactions, drug interactions, overdoses and medication errors. Identifying and notifying caregivers of potential adverse drug events (ADEs) can improve healthcare delivery (Henry et al., 2020).

Relation Type	Example
Strength-Drug	[Furosemide]$_{DRUG}$ [10 mg]$_{STRENGTH}$ IV ONCE Duration: 1 Doses.
Dosage-Drug	Patient has been switched to [lisinopril]$_{DRUG}$ 10mg [1]$_{DOSAGE}$ tablet PO QD.
Duration-Drug	Patient prescribed 1 x 20 mg [Prednisone]$_{DRUG}$ tablet daily for [5 days]$_{DURATION}$.
Frequency-Drug	Patient prescribed 1 x 20 mg [Prednisone]$_{DRUG}$ tablet [daily]$_{FREQUENCY}$ for 5 days.
Form-Drug	Patient prescribed 1 x 20 mg [Prednisone]$_{DRUG}$ [tablet]$_{FORM}$ daily for 5 days.
Route-Drug	[Furosemide]$_{DRUG}$ 10 mg [IV]$_{Route}$ ONCE Duration: 1 Doses.
Reason-Drug	Patient prescribed 1-2 325 mg / 10 mg [Norco]$_{DRUG}$ pills every 4-6 hours as needed for [pain]$_{REASON}$.
ADE-Drug	Patient is experiencing [muscle pain]$_{ADE}$, secondary to [statin]$_{DRUG}$ therapy for coronary artery disease.

Table 2: Annotations indicating relation types between pairs of entities in the n2c2 corpus.

We employ a binary classification scheme for the n2c2 corpus as previous work (Wei et al., 2020; Christopoulou et al., 2020). Under eight relation types, the training and test sets are divided into eight subsets. Each training instance contains a pair of entities which may have a possible relation

[1] https://github.com/isegura/DDICorpus

type, see Table 2. A summary of the distribution of the generated pairs of each relation type in the corpus is presented in Table 11. Table 12 shows the average sequence length of each relation type.

2.2 Supervised Machine Learning Methods

In the application of Random Forest and BiLSTM neural networks, feature engineering is a crucial stage in the preparation of data for supervised learning (Hasan et al., 2020). Pre-trained domain-specific BERT models have demonstrated remarkable success in contextualized representation learning and addressing natural language understanding tasks in biomedical and clinical domains (Alsentzer et al., 2019).

Random Forest. The implementation of RandomForestClassifier from scikit-learn library[2] is utilized in our experiments. The effectiveness and diversity of individual decision trees within the Random Forest method are directly influenced by the quality of the features employed. Instructed by Alimova and Tutubalina (2020), different features such as distance-based features, word-based features and negation words extracted from input text are prepared to train the RandomForestClassifier. Table 3 displays an example of the input features.

Sentence	Population pharmacokinetic analyses revealed that MTX, [NSAIDs]$_{GROUP}$, corticosteroids, and TNF blocking agents did not influence [abatacept]$_{DRUG}$ clearance.
token distance	10
character distance	61
punctuation distance	2
position	$[0, 2]$
bag of entities	$[0, 2, 0, 0]$
bag of words	$[0, 0, 1, ..., 1, 0, 0, 0, 0]$
negated e_1	0
negated e_2	1
hasBut	0

Table 3: Input features for the Random Forest method including distance features (token distance, character distance, punctuation distance and position), bag of words and entities, negation features[3]

.

BiLSTM networks. We implement the BiLSTM networks using PyTorch[4] and adopt the architecture proposed by Hasan et al. (2020) to tackle the RE tasks in our experiments. The input features are prepared considering syntactic and semantic information, shown in Table 4. Domain-specific pre-trained word embeddings (BioWordVec)[5] (Zhang et al., 2019) are employed as the first representations for the input sentence, embeddings for entities are obtained by averaging the word embeddings of the words within one entity.

Sentence, $\mathcal{S}$	He was administered Ibuprofen and [`Paracetamol`]$_{DRUG}$ [`500 mg`]$_{DOSAGE}$ for 3 days
$\mathbf{e_1}$	Paracetamol
$\mathbf{e_2}$	500, mg
Word Embeddings	pre-trained BioWordVec
Relative distance e_1	$[-5, -4, -3, -2, -1, 0, 1, 2, 3, 4]$
Relative distance e_2	$[-6, -5, -4, -3, -2, -1, 0, 0, 1, 2, 3]$
PoS tagging	[PRON, AUX, VERB, NOUN, CCONJ, PROPN, NUM, NOUN, ADP, NUM, NOUN]
DEP tagging	[nsubjpass, auxpass, ROOT, compound, cc, conj, nummod, dobj, case, nummod, nmod]
IOB tagging	[O, O, O, O, O, *B-DRUG*, *B-DOSAGE*, *I-DOSAGE*, O, O, O]

Table 4: Input features for the BiLSTM-based method including word embedding (BioWordVec), POS, DEP and IOB annotations at token-level.

Clinical BERT. We fine-tune the Clinical BERT model[6] (Alsentzer et al., 2019) for both corpora following the method of Wei et al. (2020), namely replacing the original entity words with their corresponding semantic types. Table 5 presents the resulting input sentences from n2c2 corpus. Each sentence from n2c2 dataset can include several entity pairs between which there can be a relation.

Original sentence	[CLS] Furosemide 10 mg IV ONCE Duration: 1 Doses
Candidate relation pairs	(1) [CLS] @Drug$ @Strength$ IV ONCE Duration: 1 Doses
	(2) [CLS] @Drug$ 10 mg @Route$ ONCE Duration: 1 Doses
	(3) [CLS] @Drug$ 10 mg IV @Frequency$ Duration: 1 Doses
	(4) [CLS] @Drug$ 10 mg IV ONCE Duration: @Dosage$ Doses

Table 5: An example of transformed samples from an original sentence from the n2c2 corpus.

2.3 Active Learning Strategies

Uncertainty-aware sampling is a common query framework in AL (Lewis and Catlett, 1994; Settles, 2009). We incorporate the principles of uncertainty and diversity in our instance selection strategies, aligning them with the imperative of querying informative instances to enhance the performance of the three distinct categories of machine learning methods for biomedical and clinical RE (Kumar and Gupta, 2020). To ensure a comprehensive evaluation of how the AL strategies impact the performance of different machine learning categories biomedical and clinical RE, we establish a random sampling strategy as a baseline and conduct experiments using Least Confidence (LC) (Settles, 2009) for all three machine learning methods. Bayesian Active Learning by Disagreement(Houlsby et al., 2011) is applied to deep learning methods, e.g.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[4] https://pytorch.org/tutorials/beginner/nlp/advanced_tutorial.html

[5] https://github.com/ncbi-nlp/BioWordVec

[6] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

BiLSTM networks and Clinical BERT. Due to the large number of parameters, training BiLSTM networks or fine-tuning BERT-based models can be both time-consuming and resource-intensive. To streamline the AL process with these methods, instances are selected in batches, namely we implement BatchBALD (Kirsch et al., 2019) instead of BALD. In the following, we describe the strategy query formations of $\phi(\cdot)$.

Least Confidence (LC). The LC strategy involves selecting the instance x from a training dataset $\mathcal{D}_{train}$ with the least confidence or most uncertain classification ($y \in Y$) in the context of probabilistic models (Settles, 2009). Given the model parameters ω to compute the most uncertainty of each sample with a prediction of $\mathbb{P}(y^*|x;\omega)$, the following formula from Settles (2009) is used:

$$\phi^{LC} = 1 - \mathbb{P}(y^*|x;\omega) \tag{1}$$

Once $\mathbb{P}(y^*|x;\omega)$ has been computed, the instance x with the highest value of ϕ^{LC} is queried. To query a batch of size $B > 1$, the top B samples $(x_1, ..., x_B)$ with the highest uncertainty values ϕ^{LC} are selected, referred to BatchLC. BatchLC combines uncertainty sampling and ranking to select a batch of unlabelled instances (Cardoso et al., 2017). If the query strategy is applied to Random Forest, they must first be modified to have probabilistic output (Lewis and Catlett, 1994).

Batch Bayesian Active Learning by Disagreement (BatchBALD). The LC strategy identifies unlabelled instances where the model expresses the lowest confidence levels, determined by its probability scores. In contrast, the BALD strategy queries unlabelled instances where a significant proportion of the model's parameter distribution samples yield incorrect predictions (Houlsby et al., 2011).

$$\phi^{BALD} = \mathbb{H}(y|x, \mathcal{D}_{train}) - E_{\omega\sim p(\omega|\mathcal{D}_{train})}[\mathbb{H}[(y|x, \omega, \mathcal{D}_{train}]] \tag{2}$$

Equation 2 of BALD explains how to balance the entropy of the model prediction (left term) and the expectation of the entropy of the model prediction over the posterior of the parameters ω (right term). It identifies instances where the model's predictions show uncertainty, when the left term is high and the right term is low, indicating disagreement between the posterior draws. BatchBALD allows for the simultaneous selection of multiple instances in a batch and strikes a balance between selecting instances with high individual uncertainty and ensuring diversity within the selected batch (Kirsch et al., 2019), see Equation 3. In BatchBALD, Monte-Carlo dropout is applied multiple times to deactivate certain neurons in the network for an input instance, resulting in multiple posterior draws. We implement the BatchBALD strategy using the BAAL[7] library (Atighehchian et al., 2022).

$$\mathrm{H}(\mathrm{y}_{1:b}|x_{1:b}, \mathcal{D}_{train}) - E_{\omega\sim p(\omega|\mathcal{D}_{train})}[\mathbb{H}[(y_{1:b}|x_{1:b}, \omega, \mathcal{D}_{train}]] \tag{3}$$

3 Experiment Setup

We employ a pool-based AL setup and word in an experimental setting, meaning that we have a training $\mathcal{D}_{train}$ and a test $\mathcal{D}_{test}$ dataset. The pseudocode of the AL experimental setting is shown in Algorithm 1.

Algorithm 1 Active Learning Loop

```
$initD_{\mathcal{L}} \leftarrow Random(\mathcal{D}_{train}, querySize)$
$\mathcal{D}_{\mathcal{U}} \leftarrow \mathcal{D}_{train} - \mathcal{D}_{\mathcal{L}}$
$annRate \leftarrow querySize/length(D_{train})$
$\omega_0 \leftarrow copyParams(\mathcal{M})$
$\mathcal{M} \leftarrow train(\mathcal{M}, \mathcal{D}_{\mathcal{L}})$
while $annRate < maxAnn$ do
    $q \leftarrow \phi(\mathcal{M}, \mathcal{D}_{\mathcal{U}}, querySize)$
    $\mathcal{D}_{\mathcal{L}} \leftarrow \mathcal{D}_{\mathcal{L}} \cup q$
    $\mathcal{D}_{\mathcal{U}} \leftarrow \mathcal{D}_{\mathcal{U}} - q$
    $annRate \leftarrow +(querySize/length(D_{train}))$
    $\mathcal{M} \leftarrow resetParams(\mathcal{M}, \omega_0)$
    $\mathcal{M} \leftarrow train(\mathcal{M}, \mathcal{D}_{\mathcal{L}})$
    $metrics \leftarrow eval(\mathcal{D}_{test}, \mathcal{M})$
```

An initial labelled dataset *init*$\mathcal{D}_{\mathcal{L}}$, consisting of 2.5% of the total training data, is randomly generated and the remaining data from the unlabelled pool $\mathcal{D}_{\mathcal{U}}$. The initial parameters of model $\mathcal{M}$ is trained on *init*$\mathcal{D}_{\mathcal{L}}$. A query strategy $\phi(\cdot)$ is applied to select another 2.5% of samples from $\mathcal{D}_{\mathcal{U}}$ based on the uncertainty estimates. New samples are added to $\mathcal{D}_{\mathcal{L}}$ and used to train $\mathcal{M}$. In each active learning step, the parameters of $\mathcal{M}$ are reset to the initial ω_0 to prevent over-fitting of the data from the first iteration (Gal et al., 2017; Hu et al., 2018). The evaluation metrics are computed on the test set $\mathcal{D}_{test}$ at the end of each step. The AL step is iteratively executed until the maximum annotation rate (*maxAnn*) is attained (Siddhant and Lipton, 2018).

[7] https://baal.readthedocs.io/en/latest/

3.1 Evaluation Metrics

Performance Measures. The common scores used to measure the models' performance on the RE over both corpora with all learning settings and sampling strategies include **Precision**, **Recall** and **F1 scores**. For the RE performance on the n2c2 corpus, we compute the binary classification results for each relation type. For the DDI corpus, we compute the 5-class (4 relation types and 1 None type) results of different relation types.

Active Learning Step Time. In the AL experiments, we evaluate the performance of three machine learning methods using up to 50% of the n2c2 and DDI training dataset respectively. At each AL step, each query strategy samples 2.5% of the data, for a total of 20 steps. We compare the performance of different AL sampling strategies, such as LC, BatchLC and BatchBALD, to a random baseline (i.e. random sampling) (Settles and Craven, 2008; Shelmanov et al., 2019; Siddhant and Lipton, 2018). Within our experimental framework, we focus on the **Step Time** taken from querying new instances to model retraining. We compare the efficiency of different AL strategies in conjunction with different machine learning methods based on the step time metrics. They provide invaluable insights into the comparative effectiveness of the strategies under investigation.

Token Annotation Rate. We omit the real-world manual annotation process in the AL experiments. However, an assumption is that longer samples generally necessitate more time for reading, analysis and annotation (Kholghi et al., 2015). Consequently, query strategies favouring the querying of lengthier samples are likely to incur higher manual annotation costs. To assess whether the process queried shorter, longer, or uniformly all lengths of samples in the unlabelled pool, we measure the number of labelled annotation units in terms of Token Annotation Rate (TAR) and Characters Annotation Rate (CAR). These metrics (Equations 4 and 5) of the annotation effort are calculated when new instances are sampled up to 50% of the training dataset.

$$TAR = \frac{no.\ of\ labelled\ tokens}{total\ no.\ of\ tokens} \tag{4}$$

$$CAR = \frac{no.\ of\ labelled\ characters}{total\ no.\ of\ characters} \tag{5}$$

4 Results and Analysis

4.1 Performance in two Corpora

Table 6 shows that both the Random Forest and Clinical BERT methods achieve better F1 scores in the AL setting using 50% of the data compared to the passive learning setting using the entire training dataset in both corpora. In the n2c2 corpus, F1 scores are consistently above 90% for the majority of relation types. Two key factors contribute to these high scores. First, relying on entity types to determine relation types simplifies the task, requiring methods to focus on relation identification rather than classification. Secondly, the distinct structural patterns associated with most relation types facilitate straightforward identification. In particular, challenges arise with more complicated types such as ADE-Drug and Reason-Drug, as highlighted in the 2018 n2c2 challenge (Henry et al., 2020). Moreover, Clinical BERT achieves a notable improvement after just a few AL steps. This improvement hints at a potential overfitting of the whole n2c2 corpus training set with Clinical BERT.

The classification of drug-drug interactions in the DDI corpus presents a more challenging task. First, the model must not only identify these interactions but also classify their types. This complexity is exacerbated by the imbalance in the relation annotations of the dataset, a factor that significantly affects the F1 scores obtained by all methods in the passive learning environment. However, the F1 scores achieved in AL setting demonstrate significant performance with considerably less annotated data.

4.2 Performance of ML Methods

In terms of performance, Clinical BERT achieves the highest F1 scores on both datasets in all settings. Notably, Random Forest emerged as the second-best method in the AL setup. However, it still presents a more challenging task of DDI corpus due to the lack of a clear text structure of the relations and the requirement to identify and classify different types of relations. Clinical BERT exhibits a remarkable improvement in their performance on the DDI corpus by utilising much fewer learning samples of the annotated data (see Table 6 and Figure 2). This demonstrates the superior ability of language models to comprehend language and generalise to various types of corpora and text-related tasks. The underperformance of Random Forest on the DDI corpus suggests that the features employed

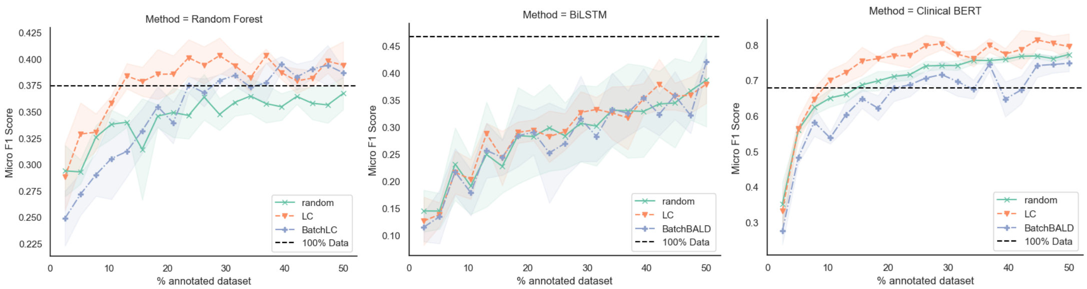
(a) Corpus = DDI

Method	**Detection**	**Effect**	**Mechanism**	**Advise**	**Int**	**Macro**	**Micro**
Random Forest	**.645 ± .01** ^{2}	**.484 ± .02** ^{2}	**.411 ± .01** ^{2}	**.464 ± .01** ^{2}	**.413 ± .03** ^{1}	**.390 ± .02** ^{2}	**.418 ± .00** ^{2}
BiLSTM	.564 ± .03 ^{4}	.445 ± .03 ^{2}	.448 ± .03 ^{4}	.488 ± .03 ^{4}	**.418** ± .04 ^{1}	.408 ± .03 ^{4}	.425 ± .02 ^{4}
Clinical BERT	**.882** ± **.00** ^{2}	**.792 ± .02** ^{2}	**.847** ± **.01** ^{2}	**.888** ± **.01** ^{1}	**.579** ± **.01** ^{2}	**.815** ± **.04** ^{2}	**.839** ± **.03** ^{2}

(b) Corpus = n2c2

Method	**Strength**	**Duration**	**Route**	**Form**	**ADE**	**Dosage**	**Reason**	**Frequency**	**Macro**	**Micro**
Random Forest	**.981 ± .00** ^{2}	**.915 ± .00** ^{2}	**.976 ± .00** ^{2}	**.988 ± .00** ^{3}	**.860 ± .00** ^{3}	**.975 ± .00** ^{2}	**.879 ± .01** ^{3}	**.963 ± .00** ^{2}	.931 ± .00 ^{3}	**.954 ± .00** ^{2}
BiLSTM	**.963 ± .00** ^{1}	**.859 ± .01** ^{2}	.947 ± .01 ^{2}	.968 ± .00 ^{4}	**.840 ± .02** ^{1}	.946 ± .00 ^{1}	**.839 ± .03** ^{2}	**.941 ± .01** ^{4}	.856 ± .04 ^{2}	.908 ± .01 ^{1}
Clinical BERT	**.992 ± .00** ^{2}	**.908 ± .01** ^{4}	**.993 ± .00** ^{2}	**.990 ± .00** ^{1}	**.888 ± .01** ^{2}	**.992 ± .00** ^{2}	**.935 ± .01** ^{2}	**.992 ± .00** ^{2}	.944 ± .00 ^{4}	**.969 ± .00** ^{2}

Table 6: F1 scores with the optimal query strategy in the AL setting, indicated by superscripts (1: Random Sampling, 2: Least Confidence, 3: BatchLC, 4: BatchBALD), are presented alongside the different machine learning methods. The annotation set is capped at a maximum of 50% of the complete training dataset. The presented F1 scores in both tables depict the mean and standard deviation of the best scores achieved for each relation type, alongside macro and micro results for the entire test set under different query strategies. Superior results, when compared to the passive learning setting with 100% training data, are highlighted in bold.

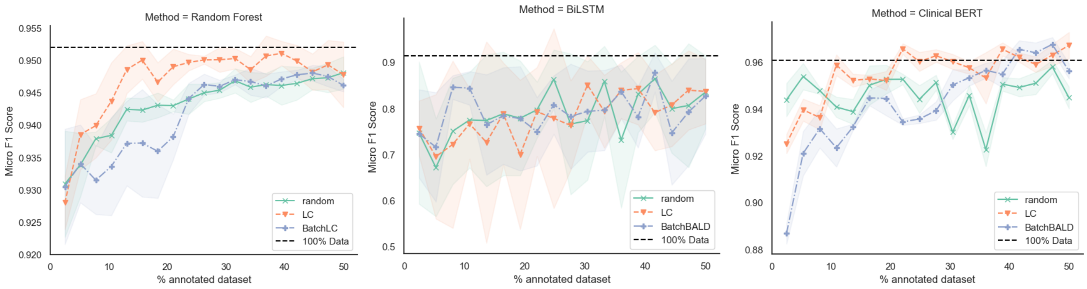


(a) Random Forest

(b) BiLSTM networks

(c) Clinical BERT

Figure 2: Micro-averaged F1 scores evolution of the different methods and query strategies during the AL process on the DDI corpus. The x-axis represents the percentage of annotated data and the y-axis represents the scores. The dashed black line indicates average performance using 100% of the data in the passive learning setting. Each line represents the average performance evolution for a query strategy. The F1 scores are computed after every AL step. The shaded area shows the standard deviation of this evolution across experiment repetitions.



(a) Random Forest

(b) BiLSTM networks

(c) Clinical BERT

Figure 3: Micro-averaged F1 scores evolution of the different methods and query strategies during the AL process on the n2c2 corpus. Line charts depicting the evolution of scores of separately trained binary models on the different n2c2 relation types are presented. The x-axis represents the percentage of annotated data and the y-axis represents the scores.

Method	Strategy	n2c2			DDI		
		Min.	Avg.	Max.	Min.	Avg.	Max.
Random Forest	random	.48 ± .02	.62 ± .02	.82 ± .05	1.08 ± .04	1.95 ± .12	3.34 ± .26
	LC	.48 ± .00	.64 ± .02	.88 ± .03	1.06 ± .06	1.97 ± .16	3.12 ± .18
	BatchLC	.66 ± .03	1.38 ± .04	1.94 ± .06	3.69 ± .19	14.10 ± .35	20.29 ± .79
BiLSTM	random	.43 ± .01	.81 ± .01	1.20 ± .02	.85 ± .27	2.79 ± .32	4.77 ± .47
	LC	.43 ± .01	.82 ± .01	1.21 ± .02	.87 ± .25	2.79 ± .33	4.88 ± .48
	BatchBALD	2.92 ± .01	3.18 ± .02	3.45 ± .03	30.17 ± 1.03	39.39 ± .87	48.48 ± 1.17
Clinical BERT	random	.72 ± .00	3.60 ± .02	6.57 ± .07	2.35 ± .02	21.64 ± .15	41.61 ± .07
	LC	.73 ± .02	3.61 ± .03	6.55 ± .05	2.35 ± .01	21.63 ± .16	41.92 ± .55
	BatchBALD	2.96 ± .14	5.82 ± .30	8.83 ± .91	12.08 ± .16	31.43 ± .36	51.64 ± .61

Table 7: Minimum, average and maximum active learning step times (in minutes). Mean and standard deviation are reported for each method and query strategy. For the n2c2 corpus, results display a weighted average across the different relation types.

Method	Strategy	n2c2		DDI	
		TAR (%)	CAR (%)	TAR (%)	CAR (%)
Random Forest	random	50.14 ± 0.17	50.16 ± 0.17	50.01 ± 0.20	50.08 ± 0.28
	LC	**47.88 ± 1.43**	**47.72 ± 1.46**	**38.11 ± 0.40**	**35.85 ± 0.29**
	BatchLC	65.94 ± 0.05	66.39 ± 0.06	45.82 ± 0.30	45.09 ± 0.70
BiLSTM	random	**49.91 ± 0.24**	**49.88 ± 0.23**	50.05 ± 0.10	50.04 ± 0.07
	LC	51.01 ± 0.59	50.93 ± 0.55	49.96 ± 0.17	49.98 ± 0.16
	BatchBALD	50.09 ± 0.08	50.07 ± 0.10	**47.66 ± 0.23**	**47.84 ± 0.20**
Clinical BERT	random	50.27 ± 0.42	50.13 ± 0.44	47.84 ± 0.81	47.59 ± 1.14
	LC	**47.91 ± 0.70**	**47.55 ± 0.28**	**36.22 ± 0.49**	**37.81 ± 0.69**
	BatchBALD	48.91 ± 0.30	48.44 ± 0.20	47.56 ± 1.15	46.55 ± 0.69

Table 8: The tar and car percentages attained after annotating 50% of instances are reported. The mean and standard deviation for each method and query strategy on both corpora are presented. Results for the n2c2 corpus are a weighted average across various relation types. The minimum tar and car values for each method on each corpus are highlighted in bold.

may struggle in capturing the specific characteristics necessary for accurate relation identification and classification. Addressing this performance gap between Random Forest and Clinical BERT would necessitate a significantly higher investment of effort in the feature engineering process for the Random Forest method.

The results of the BiLSTM networks proposed by Hasan et al. (2020) in the passive learning setting reveal its suitability for the RE in medical text. However, its performance in the AL process is sub-optimal. The method exhibits highly variable performance as illustrated in Figures 2 and 3. Although the model performance progressively improved during the AL process, neither LC nor BatchBALD demonstrates a discernible improvement over the random sampling baseline. These findings are also reflected in Table 6.

4.3 Effectiveness of AL Strategies

The analysis of the evolution of the F1 scores during the AL process of the different query strategies (Figures 2 and 3) shows that the LC strategy consistently outperformed the random baseline across the two corpora with both the Random Forest and Clinical BERT methods. Passive learning involving training a model on the entire training dataset, is used as a reference to determine the possible highest performance. Conversely, neither BatchLC nor BatchBALD consistently outperformed the random baseline across the two corpora. These batch-based strategies aim to select informative and representative batches of samples, overcoming the selection of redundant samples that simpler query strategies may exhibit. The observed inability of these batch-based query strategies to achieve significant improvements in performance prompts further investigation.

4.4 AL Step Time of ML methods

If there were no time constraints and sufficient computational resources, Clinical BERT would undoubtedly be the most appropriate method for RE tasks in medical domains. However, in an AL setting, where human annotators collaborate with the ML models, the time required for retraining and querying a new set of instances becomes an important consideration in the selection of ML methods. Table 7 shows the significant difference in the step time of different ML methods that a human expert can expect to invest in annotating a specific medical text corpus, including both the annotation itself and the waiting time for retraining and querying additional samples. For example, using the LC strategy, the Clinical BERT method takes a total of 68.48 minutes on the n2c2 corpus and 410.88 minutes on the DDI corpus. In contrast, the Random Forest method only requires 12.19 and 37.43 minutes respectively for each corpus. Consequently, this aspect may overshadow the benefits of the superior generalisation capabilities of Clinical BERT, potentially rendering this method unsuitable for an interactive learning process.

4.5 Annotation Rates

Previous studies have measured the amount of annotation effort saved by different query strategies to achieve specific performance goals through different annotation rates (Kholghi et al., 2015, 2016). The inclination of AL strategies to select longer samples from the dataset is likely attributed to their potential for exhibiting increased uncertainty (Settles, 2009). However, this practice may extend the annotation time required by human experts for thorough reading and analysis, especially if a query strategy consistently opts for longer samples. In our experimental setup, all employed strategies harnessed up to 50% of the available data. Consequently, if both TAR and CAR values remain be-low 50.00, the annotation process predominantly involves querying shorter instances (see Table 7). Conversely, if these values are above 50.00, the AL process focuses primarily on querying longer instances. This nuanced exploration highlights the dynamic relationship between query strategies, instance length and the resulting annotation effort. In particular, potential annotation savings are observed when using the LC strategy in conjunction with the Random Forest and Clinical BERT methods.

5 Related Work

Previous research has shown that traditional ML approaches yield comparable results in biomedical RE tasks with limited data instances, while deep learning models excel when more data is available (Munkhdalai et al., 2018; Xu et al., 2017; Bose et al., 2021b; Magge et al., 2018; Shelmanov et al., 2019; Christopoulou et al., 2020; Alimova and Tutubalina, 2020; Hasan et al., 2020). Previous works have also demonstrated that by annotating fewer samples selected with AL strategies, the same or even better performance can be achieved in the field of biomedical information extraction tasks (Zhang et al., 2012; Kholghi et al., 2015; Shelmanov et al., 2021; Sheng et al., 2020; Ein-Dor et al., 2020). In a more recent study, Wright et al. (2022) used a pre-trained SciBERT model for biomedical relation extraction, employing uncertainty sampling to prioritize predictions.

Siddhant and Lipton (2018) provided an empirical study of deep AL addressing multiple tasks. (Kirsch et al., 2019) proposed BatchBALD, which considered dependencies within an acquisition batch and showed increased diversity of data points and improved performance over BALD (Houlsby et al., 2011) and other methods. Zhang et al. (2012) proposed an AL framework for biomedical relation extraction, addressing key issues like query strategies, data diversity selection, and informative feature selection. The suggested query strategies include an uncertainty-based method using *Maximum Entropy* and a density-based method with K-Means clustering. Kholghi et al. (2015) empirically compared AL query strategies for clinical information extraction. They introduced a novel approach incorporating informativeness with domain knowledge, achieving equivalent performance with only 55% of the training data on the 2010 i2b2/VA concept extraction task. Chen et al. (2015) evaluated ten AL query strategies for named entity recognition (NER), finding that uncertainty-based sampling algorithms outperformed others. The varying perspectives on annotation time considerations, as seen in Chen et al. (2015) and Kholghi et al. (2015), underscore the importance of carefully selecting metrics in AL methodologies. Collectively, these studies contribute to a deeper understanding of effective AL strategies and their impact on diverse ML tasks in the medical information extraction domain.

However, a compelling need emerges for a comprehensive exploration of the inherent trade-offs in the performance of diverse ML methods. This necessity is underscored by the observed lack of attention to the critical balance between performance and the cost implications associated with various tasks in real-world applications. In response to this gap, our research aims to conduct a nuanced analysis of the broader implications, ultimately providing valuable insights to guide optimal ML methods and AL strategies within the dynamic context of interactive machine learning for medical information extraction.

6 Conclusion

Our experimental results and comparative analysis demonstrate the effectiveness of AL in optimising Clinical BERT for RE tasks in both biomedical and clinical corpora. This optimisation allows for a significant reduction in the amount of annotated data required, thereby reducing the costs associated with annotating complex medical texts. Despite the notable advantages of Random Forest, characterised by its simpler design and shorter AL step times, it requires a significant up-front investment in feature engineering. This requirement becomes particularly pronounced when dealing with data from a novel domain, thereby influencing the overall cost of the annotation process. Clinical BERT benefits from the integration of AL strategies, demonstrating improved performance with significantly reduced training data requirements. Considering the AL step time of the Random Forest method as an upper bound, future research efforts in optimising BERT-based methods for biomedical and clinical RE are imperative to address the challenges associated with increased computational time and potential inefficiencies during the interactive annotation process.

Limitation

The experiments in this work focus on biomedical and clinical corpora, which have specific linguistic nuances and subtleties inherent to the medical domain. As a result, the findings may not be universally applicable and seamlessly generalisable to other domains characterised by different terminologies, structures and linguistic patterns. Although the experiments acknowledge the potential impact of increased computational time and resource requirements, particularly in the context of interactive annotation processes, the scalability of Clinical BERT to larger datasets or real-time applications may be limited by resource constraints and may affect the efficiency of the AL process.

Acknowledgements

This research has been supported by the pAItient project (BMG, 2520DAT0P2), the Ophthalmo-AI project (BMBF, 16SV8639) and the Endowed Chair of Applied Artificial Intelligence, Oldenburg University.

References

Ilseyar Alimova and Elena Tutubalina. 2020. Multiple features for clinical relation extraction: A machine learning approach. *Journal of Biomedical Informatics*, 103:103382.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, WA Redmond, and Matthew BA McDermott. 2019. Publicly available clinical bert embeddings. *NAACL HLT 2019*, page 72.

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120.

Parmida Atighehchian, Frederic Branchaud-Charron, Jan Freyberg, Rafael Pardinas, Lorne Schell, and George Pearse. 2022. Baal, a bayesian active learning library. https://github.com/baal-org/baal/.

Priyankar Bose, Sriram Srinivasan, William C. Sleeman, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. 2021a. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18).

Priyankar Bose, Sriram Srinivasan, William C Sleeman IV, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. 2021b. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18):8319.

Thiago NC Cardoso, Rodrigo M Silva, Sérgio Canuto, Mirella M Moro, and Marcos A Gonçalves. 2017. Ranked batch–mode active learning. *Information Sciences*, 379:313–337.

Yukun Chen, Thomas A Lasko, Qiaozhu Mei, Joshua C Denny, and Hua Xu. 2015. A study of active learning methods for named entity recognition in clinical text. *Journal of Biomedical Informatics*, 58:11–18.

Fenia Christopoulou, Thy Thy Tran, Sunil Kumar Sahu, Makoto Miwa, and Sophia Ananiadou. 2020. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, 27(1):39–46.

John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–37.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning – Volume 70*, ICML'17, pages 1183–1192. JMLR.org.

Fatema Hasan, Arpita Roy, and Shimei Pan. 2020. Integrating text embedding with traditional nlp features for clinical relation extraction. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 418–425. IEEE.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *stat*, 1050:24.

Peiyun Hu, Zachary C. Lipton, Anima Anandkumar, and Deva Ramanan. 2018. Active learning with partial feedback. *ArXiv*, abs/1802.07427.

Mahnoosh Kholghi, Laurianne Sitbon, Guido Zuccon, and Anthony Nguyen. 2015. External knowledge and query strategies in active learning: A study in clinical information extraction. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 143–152.

Mahnoosh Kholghi, Laurianne Sitbon, Guido Zuccon, and Anthony Nguyen. 2016. Active learning: A step towards automating medical concept extraction. *Journal of the American Medical Informatics Association*, 23(2):289–296.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Punit Kumar and Atul Gupta. 2020. Active learning query strategies for classification, regression, and clustering: a survey. *Journal of Computer Science and Technology*, 35:913–945.

David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier.

Siting Liang, Mareike Hartmann, and Daniel Sonntag. 2023. Cross-lingual german biomedical information extraction: from zero-shot to human-in-the-loop. *arXiv e-prints*, pages arXiv–2301.

Seiji Maekawa, Dan Zhang, Hannah Kim, Sajjadur Rahman, and Estevam Hruschka. 2022. Low-resource interactive active labeling for fine-tuning language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3230–3242, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arjun Magge, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2018. Clinical ner and relation extraction using bi–char–lstms and random forest classifiers. In *International Workshop on Medication and Adverse Drug Event Detection*, pages 25–30. PMLR.

Tsendsuren Munkhdalai, Feifan Liu, Hong Yu, et al. 2018. Clinical relation extraction toward drug safety surveillance using electronic health record narratives: Classical learning versus deep learning. *JMIR public health and surveillance*, 4(2):e9361.

Burr Settles. 2009. Active learning literature survey.

Burr Settles. 2012. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079. Association for Computational Linguistics.

Artem Shelmanov, Vadim Liventsev, Danil Kireev, Nikita Khromov, Alexander Panchenko, Irina Fedulova, and Dmitry V. Dylov. 2019. Active learning with deep pre–trained models for sequence tagging of clinical and biomedical texts. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 482–489.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, and Alexander Panchenko. 2021. Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1698–1712, Online. Association for Computational Linguistics.

Ming Sheng, Jing Dong, Yong Zhang, Yuelin Bu, Anqi Li, Weihang Lin, Xin Li, and Chunxiao Xing. 2020. Ahiap: An agile medical named entity recognition and relation extraction framework based on active learning. In *Health Information Science: 9th International Conference, HIS 2020, Amsterdam, The Netherlands, October 20–23, 2020, Proceedings 9*, pages 68–75. Springer.

Aditya Siddhant and Zachary C. Lipton. 2018. Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2904–2909, Brussels, Belgium. Association for Computational Linguistics.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Zijie J Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52.

Qiang Wei, Zongcheng Ji, Yuqi Si, Jingcheng Du, Jingqi Wang, Firat Tiryaki, Stephen Wu, Cui Tao, Kirk Roberts, and Wang Qi. 2020. Relation extraction from clinical narratives using pre-trained language models. *AMIA Annual Symposium Proceedings*, 2019:1236–1245.

Dustin Wright, Anna Lisa Gentile, Noel Faux, and Kristen L Beck. 2022. Bioact: Biomedical knowledge base construction using active learning. *bioRxiv*, pages 2022–04.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

Jun Xu, Hee-Jin Lee, Zongcheng Ji, Jingqi Wang, Qiang Wei, and Hua Xu. 2017. Uth_ccb system for adverse drug reaction extraction from drug labels at tac–adr 2017. In *TAC*.

Shweta Yadav, Srivatsa Ramesh, Sriparna Saha, and Asif Ekbal. 2022. Relation extraction from biomedical and clinical text: Unified multitask learning framework. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(2):1105–1116.

Hong-Tao Zhang, Min-Lie Huang, and Xiao-Yan Zhu. 2012. A unified active learning framework for biomedical relation extraction. *Journal of Computer Science and Technology*, 27(6):1302–1313.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):52.

Deyu Zhou, Dayou Zhong, Yulan He, et al. 2014. Biomedical relation extraction: from binary to complex. *Computational and mathematical methods in medicine*, 2014.

A Statics of Datasets

Table 9 provides a breakdown of the training and test set sizes within the DDI corpus, including counts of instances annotated with various relation types. Instances labelled as *Negative* signify the absence of identified relations between entities. The data presented in Table 10 reveals the average sequence length of instances categorized by different relation types. Sequence length serves as an indicator of sentence complexity, impacting the workload associated with analyzing and annotating the sentences.

Relation	**Train**	**Test**	**Overall**
Effect	1684	360	2044
Mechanism	1312	302	1614
Advise	823	221	1044
Int	189	96	285
Positive	4008	979	4987
Negative (NO-REL)	23697	4724	28421
Overall	27705	5703	33408

Table 9: Number of annotated relations in the DDI corpus. *Positive* corresponds to the sum of *Effect*, *Mechanism*, *Advise* and *Int*

Relation	**Train**	**Test**
Effect	28.54	25.74
Mechanism	30.11	28.59
Advise	27.49	28.45
Int	37.13	35.79
Negative (NO-REL)	41.36	37.10
Overall	39.60	35.58

Table 10: Average sequence lengths (i.e. number of tokens) in the DDI corpus

Table 11 provides statics of the training and test subsets based on each relation type within the n2c2 corpus. Table 12 reveals the average sequence length of the instances containing at least one relation type.

Relation		**Train**			**Test**		**Overall**
	positive	**negative**	**total**	**positive**	**negative**	**total**	
Strength-Drug	6579	8302	14881	4237	6018	10255	25136
Duration-Drug	402	236	638	426	142	568	1206
Route-Drug	2837	3108	5945	3544	3240	6784	12729
Form-Drug	4127	1836	5963	4374	1008	5382	11345
ADE-Drug	800	367	1167	732	249	981	2148
Dosage-Drug	1528	1690	3218	2694	869	3563	6781
Reason-Drug	2987	1499	4486	3407	928	4335	8821
Frequency-Drug	3484	6456	9940	4029	5189	9218	19158
Overall	22744	23494	46238	23443	17643	41086	87324

Table 11: Number of annotated relations in the n2c2 corpus

Relation	**Train**	**Test**
Strength-Drug	26.08	37.60
Duration-Drug	27.56	25.98
Route-Drug	27.59	41.70
Form-Drug	21.82	18.38
ADE-Drug	24.93	26.73
Dosage-Drug	29.51	23.30
Reason-Drug	26.38	28.27
Frequency-Drug	31.77	41.76
Overall	27.21	34.05

Table 12: Average sequence lengths (i.e. number of tokens) in the n2c2 corpus