Optimizing Relation Extraction in Medical Texts through Active Learning:

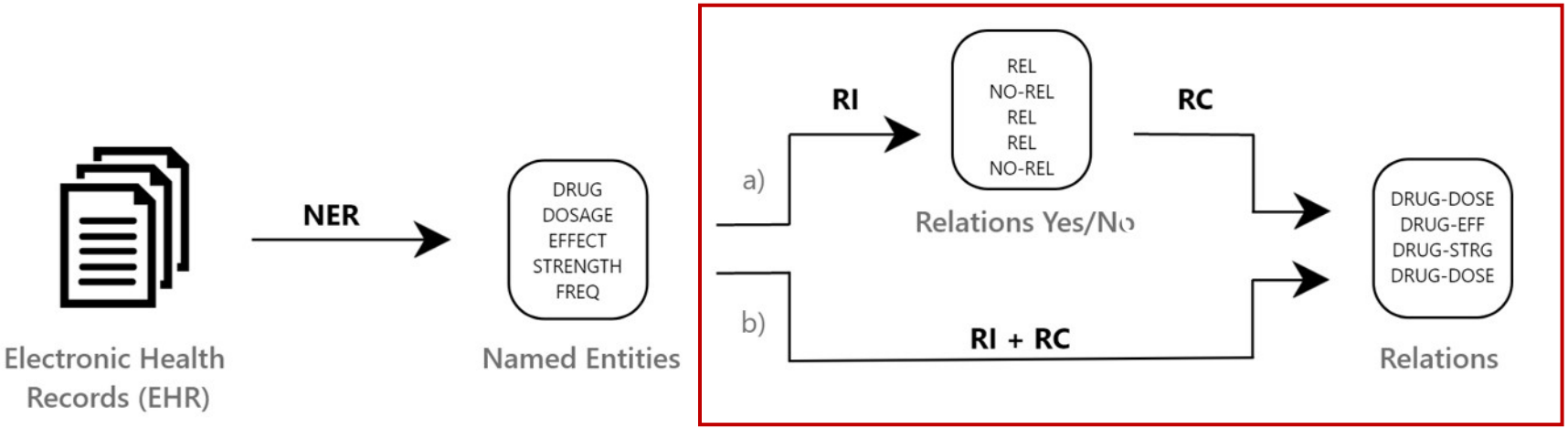
A Comparative Analysis of Trade-offs

Siting Liang[1], Pablo Valdunciel Sánchez, Daniel Sonntag[1]*
[1]German Research Center for Artificial Intelligence, Germany
*University of Oldenburg, Germany

The 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)

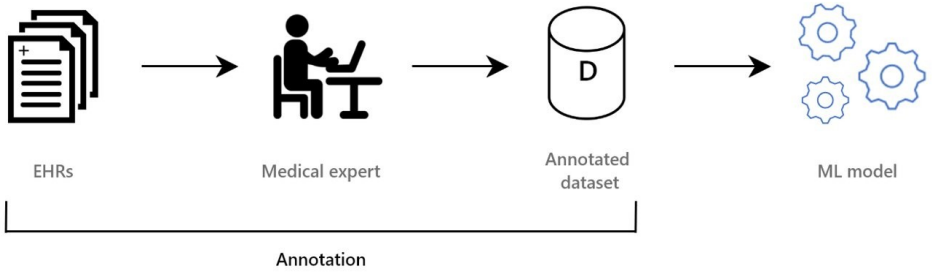
Background and Motivation

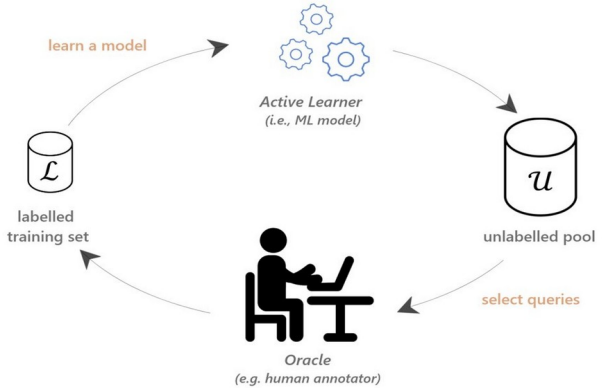
- **Clinical Information Extraction (IE)**
  - Named Entity Recognition (diagnoses, treatment, medication, disease ...)
  - Relation between entities (disease-medication)



Relation Extraction Pipeline (NER: Named-Entity Recognition; RI: Relation Identification; RC: Relation Classification)

Motivation





- Machine learning models rely heavily on the training data.
- Annotation specifications are preliminary to the setup of the clinical information system.

- Active learning advocates for a gradual labelling approach focused on the most informative instances by strategically selecting challenging or uncertain instances to minimize annotation efforts while preserving model performance.

Goal

- This work aims to assess potential variations in annotation costs for RE in medical text using various supervised learning methods with AL.
  - Domain specific language models (e.g. Clinical BERT) demonstrate exceptional performance across various tasks.
  - Active Learning (AL) with pre-trained language models can result in unacceptable waiting time for annotators.
  - Traditional machine learning (ML) models emerge as potentially more suitable choices for human-in-the-loop settings due to their shorter iteration times.

- The primary objective is to understand the data requirements of different methods for the tasks and identify resource-efficient strategies for real-world applications.

Data

- n2c2 corpus (Henry et al., 2020)
  - **Discharge summaries** from the MIMIC-III clinical care dataset (Johnson et al., 2016)
  - Extraction of medication and medication-related information from clinical texts, with focus on Adverse Drug Events (ADEs).

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

Number of annotated relations in the n2c2 corpus

- Drug-drug interaction corpus (Herrero-Zazo et al., 2013)
  - Abstracts of biomedical publications
  - Detection of pharmacological substances and drug-drug interactons in biomedical texts

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

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

Supervised Machine Learning Methods

Deutsches Forschungszentrum für Künstliche Intelligenz
German Research Center for Artificial Intelligence

Traditional ML Method

- Random Forest Classifier

Features:

Token distance, positions, bag of words, bag of entities, negation words

(Alimova and Tutubalina et al., 2020**)**

Deep Learning Neural Network

- BiLSTM-based Method

Features:

POS, DEP, entity labels, entity distance

(Hasan et al., 2020)

Language Model-Based Method

- Clinical BERT-based Method

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

(Wei et al., 2020)

AL Query Strategies

Random Sampling

Uncertainty Based

Select the instances the model is most uncertain about

Pros:
- Intuitive, easy to implement

Cons:
- Sampling bias
- Outliers

e.g.:
- **Least Confidence (LC)**

Informative & Representative

e.g.:
- Pre-clustering + Least Confidence
- **BatchLC, BatchBALD**

Experimental Setting

dfki ai Deutsches Forschungszentrum für Künstliche Intelligenz
German Research Center for Artificial Intelligence

Corpora

2013 DDI Extraction
2018 n2c2 challenge

X

Methods

Random Forest
BiLSTM
Clinical BERT

X

Query strategies

random sampling
Least Confidence (LC)
BatchLC / BatchBALD

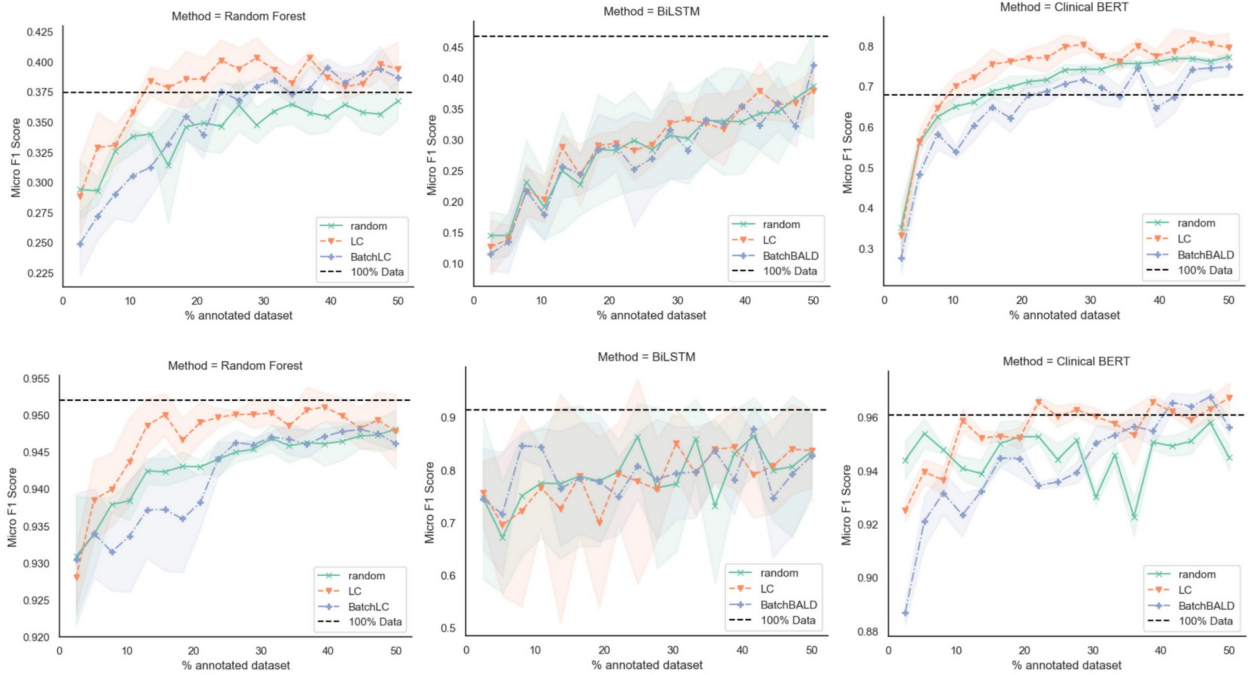
X

Evaluation Metrics

Precision
Recall
F1-score
Active Learning Step Time
Token Annotation Rate
Character Annotation Rate

- **RQ1**: How does the effectiveness of active learning in RE task in medical texts compare to that of a passive learning approach, when using less annotated data?
- **RQ2**: Which is the most cost-effective category of supervised learning methods for use in an active learning setting for the tasks?

Results



- Evaluation results of different methods and query strategies during the AL process on the DDI (above) and n2c2 corpus (below).
- The dashed black line indicates average performance using 100% of the data in the passive learning setting.
- Clinical BERT exhibits a remarkable improvement in their performance on the DDI corpus by utilising much fewer learning samples of the annotated data.

Results

(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}

- 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.
- Superior results, when compared to the passive learning setting with 100% training data, are highlighted in bold.

Results

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

- Minimum, average and maximum active learning step times (in minutes).
- Mean and standard deviation are reported for each method and query strategy.

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

- The TAR and CAR approximate the annotation time required by human experts for thorough reading and analysis.
- TAR and CAR percentages attained after annotating 50% of instances are reported.
- The minimum TAR and CAR values for each method on each corpus are highlighted in bold.

Conclusions

- Our main objective is to optimize RE in medical texts through AL while examining the trade-offs between performance and computation time.

- While Clinical BERT exhibits clear performance advantages across two different corpora, the trade-off involves longer computation times in interactive annotation processes. (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.)

- 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.

- In real-world applications, where practical feasibility and timely results are crucial, optimizing this trade-off becomes imperative. (Address the challenges associated with increased computation time and inefficiency of LM-based models during the interactive annotation process.)

Thank You !