TU WIEN Informatics

Datensatzübergreifende Erkennung medizinischer Entitäten

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Data Science

eingereicht von

Nils Kopali, Bsc
Matrikelnummer 01627943

an der Fakultät für Informatik
der Technischen Universität Wien

Betreuung: Univ.Ass. / PhD. Gábor Recski

Wien, 2. November 2024

Nils Kopali

Gábor Recski

TU WIEN Informatics

Cross-dataset medical entity recognition

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Nils Kopali, Bsc
Registration Number 01627943

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Ass. / PhD. Gábor Recski

Vienna, November 2, 2024

Nils Kopali	Gábor Recski

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Nils Kopali, Bsc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 2. November 2024

Nils Kopali

Acknowledgements

I'm deeply thankful to everyone who supported me throughout this thesis. My advisor, Gábor Recski, deserves special thanks for his guidance, patience, and invaluable insights that truly shaped this work. I'm also very grateful to my family and friends, whose encouragement kept me going through every challenge.

Kurzfassung

Diese Arbeit konzentriert sich auf die Robustheit und Generalisierbarkeit der Named Entity Recognition (NER) Modelle BioBERT [LYK+19] und KeBioLM [YLT+21] über biomedizinische Datensätze hinweg. Angesichts der wachsenden Komplexität biomedizinischer Texte ist die Anpassungsfähigkeit dieser Modelle an verschiedene Datensätze wichtig. Das Ziel der Studie ist es, diese Unterschiede zu analysieren, indem die Leistung der Modelle auf spezifischen Datensätzen bewertet wird, wobei der Schwerpunkt darauf liegt, wie ungesehene Entitäten und Annotationsinkonsistenzen ihre Genauigkeit und Generalisierungsfähigkeiten beeinflussen.

Wir werden die Präzision, den Recall und die F1-Werte von BioBERT und KeBioLM durch systematische Tests mit zwei biomedizinischen NER-Datensätzen, BC5CDR [LSJ+16] und NCBI [DLL14], untersuchen. Beide Modelle schneiden innerhalb eines Datensatzes gut ab; ihre Genauigkeit nimmt jedoch in datensatzübergreifenden Szenarien deutlich ab, was die Schwierigkeiten bei der Generalisierung auf neue, ungesehene Daten verdeutlicht.

Die Ergebnisse zeigen den Bedarf an anpassungsfähigeren NER-Systemen, die mit der dynamischen und vielfältigen Natur biomedizinischer Texte umgehen können. Die für die Analyse in dieser Arbeit verwendeten Skripte sind im GitHub-Repository[1] verfügbar.

[1] Nils Kopali, *GitHub Repository for NER Analysis*, 2024. Available at: `https://github.com/nkopali/NER-cross-dataset` (Last accessed: October 31, 2024).

Abstract

This thesis focuses on the robustness and generalizability of Named Entity Recognition (NER) models BioBERT [LYK+19] and KeBioLM [YLT+21] across biomedical datasets. With the growing complexity of biomedical texts, these models' adaptability to different datasets is important. The study's goal is to analyze these differences by assessing the model performance on specific datasets, focusing on how unseen entities and annotation inconsistencies affect their accuracy and generalization capabilities.

We will investigate BioBERT and KeBioLM precision, recall, and F1 scores using systematic tests with two benchmark biomedical NER datasets, BC5CDR [LSJ+16] and NCBI [DLL14]. Both models perform well within a dataset; however, their accuracy drops significantly in cross-dataset scenarios, demonstrating the difficulties in generalizing to new, unseen data.

The findings indicate the need for more adaptable NER systems that can handle the dynamic and diverse nature of biomedical texts. The scripts used for analysis in this thesis are available in the GitHub repository[2].

[2]Nils Kopali, *GitHub Repository for NER Analysis*, 2024. Available at: `https://github.com/nkopali/NER-cross-dataset` (Last accessed: October 31, 2024).

Contents

CHAPTER 1

Introduction

1.1 Background and Motivation

Named Entity Recognition (NER) [NS07] is an important task in biomedical text mining, which identifies and classifies entities in a sentence such as diseases, medications, and other relevant terms within unstructured text. Extracting these entities is important for different applications in biomedical research. Although there have been advances in NER models, particularly those fine-tuned for biomedical contexts like BioBERT [LYK+19] and KeBioLM [YLT+21], their performance often decreases when applied across different datasets. This reduction in accuracy and generalizability is very important, as biomedical texts are diverse and continuously evolving.

The motivation for this paper comes from the need to understand and improve the robustness of NER models in cross-dataset scenarios. We aim to find the factors contributing to this performance drop and to propose strategies for improving model generalization across diverse biomedical datasets. Further details and code for this paper can be found in the GitHub repository[1].

1.2 Objectives of the Study

1. **To evaluate the impact of cross-dataset training and testing on the performance of BioBERT and KeBioLM models.** This involves assessing their precision, recall, and overall accuracy by using one dataset for training and the other for testing and vice-versa.

[1] Nils Kopali, *GitHub Repository for NER Analysis*, 2024. Available at: `https://github.com/nkopali/NER-cross-dataset` (Last accessed: October 13, 2024).

2. **To investigate the influence of unseen entities on model performance.** By removing sentences whose entities were not present during training, we aim to determine the generalizability of the models in recognizing previously seen biomedical terms.

3. **To analyze the effect of annotation artifacts on NER performance**. This objective focuses on identifying inconsistencies in dataset annotations that could affect model evaluation.

1.3 Literature Review

Named Entity Recognition (NER) is a subtask of information extraction that locates and classifies entities in text into predefined categories such as the names of persons, organizations, locations etc. In the biomedical domain, NER is particularly important as it involves the identification of entities such as diseases, drugs and genes from scientific literature and other unstructured biomedical texts. The constantly evolving vocabulary, as well as the need for high precision and recall due to the sensitive nature of medical information, contribute to the unique challenges posed by biomedical natural language processing.

A paper by Liu et al. (2021) [LXY+21] looks at domain adaptation for NER applications, emphasizing the importance of domain-adaptive pre-training (DAPT) as an essential part of their methodology. Domain-Adaptive Pre-Training (DAPT) is a method designed to enhance the performance of language models on specific tasks by pre-training them on a carefully selected subset of a larger, domain specific corpus. In DAPT, instead of training a model on all available data within a domain, the process focuses on the most relevant sections that are rich in domain-specific terms and contexts. This approach involves analyzing the larger corpus to identify and extract segments that contain a high density of relevant entities and terms, thereby ensuring that the language model is exposed to the most relevant language features and contextual nuances. Similarly, our models, KebioLM and BioBERT, have been trained on large biomedical corpora to optimize their performance in biomedical domain-specific tasks.

In a paper [K+21] focusing on the generalizability of NLP models across medical specialties, the researchers examined how well SciBERT, a variant of BERT pre-trained on scientific texts, could classify diagnosis sentiment across different medical domains using the MIMIC-III dataset. The paper revealed significant problems in model generalization, especially when models trained on one specialty's data were evaluated on another. The researchers discovered that model performance got worse when the overlap between training and test specialties reduced, indicating the challenge of cross-specialty generalization. However, they also demonstrated that increasing the training data by including several specialties improved the model's ability to generalize to previously unknown specialties.

In this research paper [KK22] the researchers present a thorough examination of the generalization performance of BioNER models. The authors analyze how these models

handle unseen biomedical entities, distinguishing between three core capabilities: memorization (how well does the model identify entities that were seen during training), synonym generalization (how well does the model identify synonyms, like Motrin and Ibuprofen, which are the same concept), and concept generalization (how well does the model identify new biomedical concepts like COVID-19). They demonstrate that, despite performing well on standardized benchmarks, BioNER models struggle significantly when generalizing to new synonyms and novel concepts. This highlights that these models may be overestimated, suggesting that benchmark metrics don't always reflect how well a model will actually perform in real-world situations.

1.4 Overview of NER Models

The merging of deep learning with advanced pre-trained language models, particularly those based on transformer architectures, has transformed the field of Named Entity Recognition (NER). This section goes into how these models, like BERT [DCLT18] and its specific adaptations, have improved the capabilities of NER systems, such as those used in biomedicine.

1.4.1 BERT

BERT is a pre-trained language model that utilizes the transformer architecture to achieve good performance on a variety of NLP tasks, including Named Entity Recognition (NER). Unlike traditional models that read text sequentially, BERT processes words in relation to all other words in a sentence simultaneously, a mechanism known as "bidirectional" processing. This approach allows the model to capture the contextual meanings of words.

BERT is pre-trained on a large corpus of text from the internet, which includes tasks like predicting missing words in sentences. This pre-training serves as a foundation that can be fine-tuned with additional training on a smaller, task-specific dataset, such as biomedical texts for NER.

1.4.2 BioBERT

BioBERT is a variant of BERT. The BioBERT model is initialized with weights from BERT. BioBERT is then further pre-trained on large biomedical corpora, including PubMed abstracts and PMC full-text articles, to adapt it specifically for biomedical text mining tasks. This additional pre-training step allows BioBERT to better understand the complex terminology and context found in biomedical literature.

1.4.3 KeBioLM

KeBioLM is similar to BioBERT but it also integrates some external biomedical knowledge during training. The model leverages the Unified Medical Language System (UMLS) to enhance entity recognition performance. Specifically, KeBioLM incorporates entities from PubMed abstracts, linking them to UMLS concepts.

1.5 Description of Datasets

We utilize two biomedical NER datasets: BC5CDR and NCBI. Both datasets were pre-partitioned into train, test and dev sets. KebioLM was fine-tuned for 60 epochs with a learning rate 1e-5 and batch size of 8 for both datasets and for BioBERT a learning rate of 5e-5, 3e-5 or 1e-5 was selected with a batch size ranging from 10, 16, 32 or 64, as for the epochs it is not clearly stated but it is mentioned more than 20. Table 2 in [YLT+21] shows that the KeBioLM model achieved an F1-score of 86.1 for BC5CDR and 89.1 for NCBI and on Table 6 from BioBERT [LYK+19] an F1-score of 86.47 for BC5CDR and 88.22 for NCBI.

1.5.1 Dataset 1: BC5CDR

The BC5CDR corpus consists of 1,500 PubMed articles, which include annotations for 4,409 chemicals, 5,818 diseases, and 3,116 chemical-disease interactions. It is widely used for evaluating biomedical NER models due to its high-quality annotations and relevance to real-world biomedical text mining tasks.

1.5.2 Dataset 2: NCBI

The NCBI Disease Corpus is another benchmark dataset comprising PubMed abstracts annotated for disease names. It includes 793 PubMed abstracts with 6,892 disease mentions.

The dataset statistics, as shown in Table 1.1, indicate that BC5CDR and NCBI differ significantly in terms of the number of sentences, tokens, and entities. NCBI has 11,249 entities entities in the training set compared to BC5CDR's 7,100. However, BC5CDR has more total sentences in both the test and dev sets, with 4,812 and 4,602 sentences respectively, compared to NCBI's 942 and 923. When comparing the unique entities between the datasets, as shown in Table 1.2, there is a notable overlap of entities, with 363 intersecting entities in the training sets, 159 in the test sets, and 149 in the dev sets. BC5CDR contains more unique entities in both the test and dev sets compared to NCBI, suggesting greater entity variability in BC5CDR's test and dev data.

Dataset	Set	Total Sentences	Total Tokens	Total Entities (B/I)	Unique Entities
BC5CDR	Train	4582	118170	7100	1462
	Test	4812	124750	7161	1371
	Dev	4602	117453	6969	1356
NCBI	Train	5432	135701	11249	1509
	Test	942	24497	2047	500
	Dev	923	23969	1877	416

Table 1.1: NER Dataset Statistics for BC5CDR and NCBI

Entity Comparison	Train Set	Test Set	Dev Set
Intersecting Entities	363	159	149
Unique Entities in BC5CDR	1099	1212	1207
Unique Entities in NCBI	1146	341	267

Table 1.2: Comparison of Entities between BC5CDR and NCBI Train, Test, and Dev Sets

CHAPTER 2

Methodology

2.1 Research Design

The methodology used in this study will assess the robustness and generalizability of Named Entity Recognition (NER) models, specifically BioBERT and KeBioLM, across the biomedical datasets. This chapter describes the experimental setup, training and evaluation procedures, and methodologies for analyzing the performance of these models.

2.2 Experimental Setup

The experimental setup involves the following steps:

1. **Model Training**: Both BioBERT and KeBioLM models are evaluated on each dataset individually to establish baseline performance. BioBERT was trained for 5 epochs with a maximum sequence length of 128 and a batch size of 32, using a seed of 1 and KeBioLM was trained for 5 epochs with a maximum sequence length of 512 and a batch size of 8, also with a seed of 1.
2. **Cross-Dataset Evaluation**: To assess the models' generalization capabilities, they are tested on a dataset other than the one on which they were trained. This cross-dataset evaluation is important for assessing how well the models can handle unseen data and diverse biomedical texts.
3. **Performance Metrics**: The models' performance is measured using standard NER metrics, including Precision, Recall, and F1 score. These metrics give a thorough picture of the models' ability to correctly recognize entities across datasets.
4. **Entity Filtering**: This involves filtering entities in the validation sets to isolate unseen entities from the training set. This allows for a focused analysis of the models' generalizability in recognizing seen biomedical terms.

CHAPTER 3

Results

3.1 Analysis of Cross-Dataset Performance

Our research focused only on disease entities using the BioBERT and KeBioLM models within the BC5CDR and NCBI datasets. This specific analysis helped us explore the models' robustness and generalizability in recognizing diseases, based on tests conducted on the designated test sets for each dataset.

3.1.1 BioBERT Performance

When BioBERT is trained and tested on the same dataset, it demonstrates strong performance metrics. For instance, training and testing on BC5CDR yields an F1 score of 83.6, with a recall of 86.2 and precision of 81.1. Similarly, training and testing on NCBI produces an even higher F1 score of 86.9, recall of 88.2, and precision of 85.6. These results indicate that BioBERT is highly effective at capturing and classifying biomedical entities within the same dataset context.

However, the performance drops considerably in cross-dataset evaluations. Training on BC5CDR and testing on NCBI results in an F1 score of 66.8, recall of 61.4, and precision of 73.3. Conversely, training on NCBI and testing on BC5CDR gives an F1 score of 66.5, recall of 64.7, and precision of 68.4. This decline highlights the challenge BioBERT faces in generalizing across different datasets. The reduction in recall, in particular, suggests difficulties in identifying entities that were not part of the training data, likely due to differences in entity distributions between the datasets.

3.1.2 KeBioLM Performance

KeBioLM shows a similar pattern but with slightly better cross-dataset performance compared to BioBERT. When trained and tested on BC5CDR, KeBioLM achieves an

F1 score of 84.4, recall of 86.6, and precision of 82.2. For NCBI, the scores are F1 of 87.2, recall of 89.0, and precision of 85.5, indicating robust performance within the same dataset.

In cross-dataset scenarios, KeBioLM maintains a relative edge over BioBERT. Training on BC5CDR and testing on NCBI results in an F1 score of 71.3, recall of 68.1, and precision of 75.0. Training on NCBI and testing on BC5CDR yields an F1 score of 67.4, recall of 65.4, and precision of 69.6. The higher F1 scores compared to BioBERT suggest that KeBioLM's integration of external biomedical knowledge may enhance its ability to generalize across different datasets, although still with noticeable performance drops.

Model	Training Dataset	F1 Score	Recall	Precision
BioBERT	BC5CDR	83.6	86.2	81.1
	NCBI	66.5	64.7	68.4
KeBioLM	BC5CDR	84.4	86.6	82.2
	NCBI	67.4	65.4	69.6

Table 3.1: Performance Metrics for BioBERT and KeBioLM on BC5CDR Test Set

Model	Training Dataset	F1 Score	Recall	Precision
BioBERT	BC5CDR	66.8	61.4	73.3
	NCBI	86.9	88.2	85.6
KeBioLM	BC5CDR	71.3	68.1	75.0
	NCBI	87.2	89.0	85.5

Table 3.2: Performance Metrics for BioBERT and KeBioLM on NCBI Test Set

The significant drop in performance for both models in cross-dataset evaluations highlights the problems of biomedical NER, specifically the differences between the BC5CDR and NCBI datasets, such as entity types and annotation guidelines. These inconsistencies have a considerable influence on the models' ability to generalize, as shown by the decline in measures such as the F1 score. This suggests that both BioBERT and KeBioLM are very sensitive to specific training data and less flexible to new, previously unseen data. The drop in recall more than precision suggests that the models are less capable of identifying entities they have not been trained on.

The performance metrics are summarized in Tables 3.1 and 3.2.

3.2 Merging Training Datasets

In order to explore whether combining the BC5CDR and NCBI training datasets would improve performance, we conducted an experiment where both datasets were merged into a single training set. This combined dataset was then tested on the individual test sets (BC5CDR and NCBI) to assess the potential benefits of using a larger, more diverse training dataset. The hypothesis was that merging the datasets would provide the model

with a broader context of named entities, which might help improve generalization across the datasets.

Model	Test Dataset	F1 Score (%)	Recall (%)	Precision (%)
BioBERT	BC5CDR	83.7	85.6	81.8
	NCBI	87.3	89.1	85.5
KeBioLM	BC5CDR	85.2	87.3	83.2
	NCBI	88.5	90.7	86.4

Table 3.3: Performance Metrics of BioBERT and KeBioLM for Merged Training Dataset

These results in Table 3.3 indicate that merging the BC5CDR and NCBI training datasets have a slight improvement on both test sets. The increase in F1 scores for both BC5CDR and NCBI suggests that the model could better generalize across entities, possibly due to the enriched training data that included a wider variety of named entities from both datasets.

3.3 Analysis of Influence of Unseen Entities on Model Performance

To investigate the adaptability and flexibility of the BioBERT and KeBioLM models in recognizing new, previously unseen biomedical terms, we evaluate their performance on validation sets before and after filtering out entities present in the training sets. This analysis provides insights into the models' ability to handle unseen entities, a crucial aspect for their generalizability and practical application in biomedical text mining. As shown in Table 3.4 below, the sample sizes of the NCBI and BC5CDR datasets decrease after applying the filtering.

Filtering Scenario	Original Samples	Filtered Samples
NCBI on BC5CDR	923	508
NCBI on NCBI	923	792
BC5CDR on NCBI	4,602	2,296
BC5CDR on BC5CDR	4,602	3,582

Table 3.4: Sample Counts Before and After Filtering

3.3.1 Models Performance

Both BioBERT and KeBioLM demonstrated significant performance improvements after filtering Tables 3.5 and 3.6. The filtering process likely enhanced the models' ability to focus on familiar entities, resulting in better extraction of relevant biomedical terms. However, there was a noticeable performance drop when both models, trained on the NCBI dataset, were validated on the BC5CDR dataset. This decline can be attributed to the difference in the number of PubMed articles in each dataset: BC5CDR contains

1,500 articles, while NCBI is built from only 793. The larger and more diverse BC5CDR dataset likely includes a broader range of entities, making it more challenging for models trained on the smaller, less varied NCBI dataset to generalize effectively to this broader context.

Model	Training Dataset	Filtering	F1 Score (%)	Recall (%)	Precision (%)
BioBERT	BC5CDR	Before	83.6	86.2	81.1
	BC5CDR	After	86.7	90.0	83.3
	NCBI	Before	66.5	64.7	68.4
	NCBI	After	61.1	69.5	54.6
KeBioLM	BC5CDR	Before	84.4	86.6	82.2
	BC5CDR	After	89.3	93.2	85.6
	NCBI	Before	67.4	65.4	69.6
	NCBI	After	55.1	62.9	49.0

Table 3.5: Performance Metrics for BioBERT and KeBioLM on BC5CDR Validation Set

Model	Training Dataset	Filtering	F1 Score (%)	Recall (%)	Precision (%)
BioBERT	BC5CDR	Before	66.8	61.4	73.3
	BC5CDR	After	86.1	90.9	81.8
	NCBI	Before	86.9	88.2	85.6
	NCBI	After	89.0	92.6	85.7
KeBioLM	BC5CDR	Before	71.3	68.1	75.0
	BC5CDR	After	82.8	87.8	78.3
	NCBI	Before	87.2	89.0	85.5
	NCBI	After	89.9	93.4	86.6

Table 3.6: Performance Metrics for BioBERT and KeBioLM on NCBI Validation Set

3.4 Error Analysis

In this section, we analyze the errors observed in our validation set to better understand the performance of our model and identify areas for improvement. We took the predictions done on the NCBI validation set with training set BC5CDR and NCBI for BioBERT model.

3.4.1 Sample Reduction through Filtering

During the filtering process, we removed sentences whose entities do not appear in the training set, resulting in a significant reduction in the number of samples:

- **Filtering NCBI on BC5CDR**: The number of samples reduced from 923 to 508.
- **Filtering NCBI on NCBI**: The number of samples reduced from 923 to 792.

3.4.2 Frequent Errors: The Case of '-'

One interesting observation is related to the entity `"-"`. This entity is mispredicted 3 times in the BC5 dataset, which had a total of 30 errors, and 9 times in the NCBI dataset, which had a total of 85 errors. To quantify the impact of these errors, we calculate the error rates as follows:

- For BC5CDR: $\left(\frac{3}{30}\right) \times 100 = 10\%$
- For NCBI: $\left(\frac{9}{85}\right) \times 100 = 10.6\%$

Therefore, the `"-"` entity accounts for approximately 10% of the validation set errors, indicating a notable area of concern.

Column	Value
Dataset	NCBI
Sentence	In 22 of the 43 **non - papillary renal cell carcinomas** , abnormally migrating DNA bands were detected by SSCP and / or HD analysis .
Gold	non - papillary renal cell carcinomas
Predicted	papillary renal cell carcinomas
Dataset	NCBI
Sentence	**Atm - deficient** thymocytes undergo spontaneous apoptosis in vitro significantly more than controls .
Gold	
Predicted	Atm - deficient

Table 3.7: Examples from the case of '-'

3.4.3 Frequent Errors: The Case of 'Tumor'

Another frequent error in the BC5CDR dataset involves the word `"tumor"`. Despite being consistently annotated as `'B'` (beginning of an entity) 16 times and `'I'` (inside an entity) 4 times in the training set, the model frequently predicts it as `'O'` (outside of an entity). This suggests that the model struggles to recognize `"tumor"` as an entity.

Additionally, it appears that `"tumor"` is annotated incorrectly in 2 instances within the BC5CDR validation dataset and in 4 instances within the NCBI validation dataset. These incorrect annotations label `"tumor"` as `'O'` instead of `'B'`. Correcting these annotations would likely improve the model's performance, potentially resulting in 2 additional correct predictions for BC5CDR and 4 additional correct predictions for NCBI.

Column	Value
Dataset	NCBI
Sentence	These data functionally define a novel genetic locus , designated PAC1 , for **prostate adenocarcinoma** 1 , involved in **tumor** suppression of human prostate carcinoma and furthermore strongly suggest that the cell death pathway can be functionally restored in **prostatic adenocarcinoma** . .
Gold	, , prostatic adenocarcinoma
Predicted	prostate adenocarcinoma , tumor, prostatic adenocarcinoma
Dataset	NCBI
Sentence	BACKGROUND & AIMS The chromosome region 18q21 has been shown to be frequently deleted in **colorectal cancers** , and such frequent allelic loss is a hallmark of the presence of a **tumor** - suppressor gene .
Gold	colorectal cancers
Predicted	colorectal cancers, tumor

Table 3.8: Examples from the case of 'Tumor'

3.4.4 Frequent Errors in NCBI: The Case of 'VHL'

Another specific error in the NCBI dataset involves the word `"VHL"`. This term is a disease and is consistently annotated as `'B'` or `'I'` in the training set. However, in the validation dataset, it was annotated as `'O'` incorrectly 6 out of 8 times. These misannotations contribute significantly to the errors in the NCBI dataset.

Correcting these annotations would improve the model's performance, potentially resulting in a more accurate identification of `"VHL"` as an entity in the validation set.

Column	Value
Dataset	NCBI
Sentence	We have therefore assessed the effect of the **VHL** gene product on VEGF expression .
Gold	
Predicted	VHL
Dataset	NCBI
Sentence	wt - **VHL** protein inhibited VEGF promoter activity in a dose - dependent manner up to 5 - to 10 - fold .
Gold	
Predicted	VHL

Table 3.9: Examples from the case of 'VHL'

3.4.5 Impact of Correcting Errors

If we correct the identified annotation errors, the model's performance could improve significantly. Here are the potential improvements:

BC5 Dataset:

- Initial Total Errors: 30
- Errors Corrected: 3 (for "-") + 2 (for "tumor") = 5
- New Total Errors: 30 - 5 = 25
- Old Error Rate: $\left(\frac{30}{508}\right) \times 100 \approx 5.9\%$
- New Error Rate: $\left(\frac{25}{508}\right) \times 100 \approx 4.9\%$

NCBI Dataset:

- Initial Total Errors: 85
- Errors Corrected: 9 (for "-") + 4 (for "tumor") + 6 (for "VHL") = 19
- New Total Errors: 85 - 19 = 66
- Old Error Rate: $\left(\frac{85}{792}\right) \times 100 \approx 10.7\%$
- New Error Rate: $\left(\frac{66}{792}\right) \times 100 \approx 8.3\%$

3.5 Error Patterns

In order to analyze the syntactic structure of the errors identified in our Named Entity Recognition (NER) models, we utilized the Stanza library, a state-of-the-art NLP toolkit that provides robust syntactic analysis. We were able to categorize the errors into three main classes: false positives, false negatives, and overlaps. A false positive occurs when the model mistakenly identifies a non-entity as an entity, while a false negative is the opposite - failing to recognize an actual entity in the text. The third class, overlap, happens when the gold standard and predicted entities overlap, yet they do not represent the exact same entity class.

3.5.1 False Positives

Through the examination of false positives generated by our Named Entity Recognition (NER), we discovered various part-of-speech (POS) patterns contributing to recognition errors. Our analysis Table. 3.11 revealed that nouns ('NOUN') accounted for the majority

of false positives, occurring 52 times, followed closely by adjectives ('ADJ') at 24 instances. In the table Table. 3.10 we can see some of these examples:

Column	Value
Dataset	NCBI
Sentence	Furthermore , the complemented hybrids undergo programmed cell **death** in vitro via a mechanism that does not require nuclear localization of p53 .
Gold	
Predicted	death
Dataset	NCBI
Sentence	Mice doubly null for atm and p53 exhibited a dramatic acceleration of **tumour** formation relative to singly null mice , indicating that both genes collaborate in a significant manner to prevent **tumorigenesis** .
Gold	tumor,
Predicted	tumor, tumorigenesis
Dataset	NCBI
Sentence	In yeast , mutations in several genes , including RTH and MSH3 , cause **microsatellite** instability .
Gold	
Predicted	microsatellite
Dataset	NCBI
Sentence	In the other **Irish** family , exons 7 and 8 failed to amplify and they were shown to be deleted .
Gold	
Predicted	Irish

Table 3.10: Examples from False positives

Pattern	Frequency
NOUN	52
ADJ	24
PROPN	19
PUNCT	7
VERB	3
SYM	2
NUM	2
ADV	1
INTJ	1
CCONJ	1

Table 3.11: False positives POS Patterns and Their Frequencies

3.5.2 False Negatives

Our analysis Table. 3.13 of the false negatives revealed several recurring part-of-speech (POS) patterns that contributed to these errors. The most frequent POS tag associated with false negatives was nouns ('NOUN'), which appeared 18 times. Adjectives ('ADJ') were the second most frequent, appearing 16 times,

Below are some specific examples Table. 3.12 of false negatives:

Column	Value
Dataset	NCBI
Sentence	2 , a region proposed to contain **tumor** suppressor gene (s) , is mutated at high frequency in human **breast cancer** .
Gold	tumor , breast cancer
Predicted	, breast cancer
Dataset	NCBI
Sentence	In patients with **stage III disease** , the respective survival rates were 59 .
Gold	stage III disease
Predicted	stage III
Dataset	NCBI
Sentence	Individuals who have rare alleles of the VNTR have an increased risk of certain types of **cancers** , including **breast cancer** (2 - 4) .
Gold	cancers, breast cancer
Predicted	breast cancer

Table 3.12: Examples from False negatives

Pattern	Frequency
NOUN	18
ADJ	16
PROPN	8
PUNCT	5
ADV	4
VERB	2
NUM	1
CCONJ	1

Table 3.13: False negatives POS Patterns and Their Frequencies

3.5.3 Overlap

Overlap errors occur when the predicted entities and the gold standard entities intersect, but do not match exactly in terms of the entity class. These errors are particularly interesting because they indicate partial recognition by the model, where it correctly

identifies some portion of the entity but fails to capture it entirely or assigns the wrong entity type.

In our analysis Table. 3.15, we noticed that the overlaps primarily involved nouns ('NOUN'), adjectives ('ADJ'), and proper nouns ('PROPN'). The most frequent POS tag associated with overlap errors was 'NOUN', occurring 23 times. This suggests that the model often partially recognizes noun-based entities but either misses the complete span or misclassifies them. Adjectives ('ADJ') were the second most common, appearing 8 times, indicating that descriptive words modifying the entities were sometimes included or excluded incorrectly.

The following examples Table. 3.14 show the overlap errors, the entities are marked with * if they overlap:

Column	Value
Dataset	NCBI
Sentence	No association was found between the presence of bilateral **breast cancer** or the number of **breast cancers** in a family and the detection of a BRCA1 mutation, or between the position of the mutation in the BRCA1 gene and the presence of ovarian cancer in a family.
Gold	breast cancer, breast cancer
Predicted	breast* cancer, breast cancer
Dataset	NCBI
Sentence	In **stage II colorectal carcinomas**, the absence of DCC identifies a subgroup of patients with lesions that behave like **stage III cancers**.
Gold	stage II colorectal carcinomas, stage III cancers
Predicted	stage II colorectal* carcinomas, stage III cancers

Table 3.14: Examples from Overlap cases

Pattern	Frequency
NOUN	23
ADJ	8
PROPN	6

Table 3.15: Overlap POS Patterns and Their Frequencies

3.5.4 Patterns in Errors

During our analysis of the errors in our Named Entity Recognition (NER) models, we discovered distinct patterns in error formation. Previously, we encountered errors involving individual entities, in which the error was limited to a single entity and did not affect adjacent ones.

The table below presents the patterns which are identified by sequences of POS tags that come one after the other, often forming the core structure of the recognized entities.

Pattern	Frequency
NOUN	29
PROPN	24
ADJ NOUN	19
ADJ	13
NOUN NOUN	7
NOUN PUNCT ADJ	4
ADV	3
PROPN NOUN	2
NOUN NUM NOUN	2
ADJ NOUN NOUN	2
NOUN SYM NOUN NOUN	2
NOUN PUNCT VERB	2
ADV PUNCT NOUN	2
ADJ NOUN ADJ	1
CCONJ	1
VERB PROPN	1
ADJ ADJ ADJ	1
ADJ CCONJ	1
INTJ	1
NOUN NUM PROPN	1
PUNCT VERB	1
NOUN PUNCT ADJ NOUN	1
ADJ ADJ NOUN	1
NOUN PUNCT NOUN ADJ NOUN	1
PROPN PUNCT	1
VERB NOUN	1
ADJ PROPN	1

Table 3.16: POS Patterns and Their Frequencies

As seen in Table 3.16, the most frequent patterns include single POS tags like ‘NOUN‘ (29 occurrences) and ‘PROPN‘ (24 occurrences), followed by combinations such as ‘ADJ NOUN‘ (19 occurrences). These patterns reflect the typical structure of named entities, which often consist of nouns and adjectives.

ADJ NOUN Pattern

One of the most notable patterns identified is the ‘ADJ NOUN‘ sequence, which occurred 19 times in our analysis. This pattern typically represents descriptive entities where an adjective and a noun together, forms a phrase that should be recognized as an entity. For instance:

Column	Value
Dataset	NCBI
Sentence	In preliminary screens , mutations of PTEN were detected in 31 % (13 / 42) of **glioblastoma** cell lines and xenografts , 100 % (4 / 4) of prostate cancer cell lines , 6 % (4 / 65) of breast cancer cell lines and xenografts , and 17 % (3 / 18) of **primary glioblastomas** .
Gold	glioblastoma, primary glioblastomas
Predicted	glioblastoma, glioblastomas
Pattern	primary=ADJ, glioblastomas=NOUN
Dataset	NCBI
Sentence	To identify possible features of the BRCA1 genomic region that may contribute to **chromosomal instability** as well as potential transcriptional regulatory elements , a 117 , 143 - bp DNA sequence encompassing BRCA1 was obtained by random sequencing of four cosmids identified from a human chromosome 17 specific library .
Gold	
Predicted	chromosomal instability
Pattern	chromosomal=ADJ, instability=NOUN
Dataset	NCBI
Sentence	Inherited mutant alleles of **familial tumour** suppressor genes predispose individuals to particular types of **cancer**.
Gold	tumour, cancer
Predicted	familial tumour, cancer
Pattern	familial=ADJ, tumour=NOUN

Table 3.17: Examples from the ADJ NOUN pattern

These examples highlight the model's challenges in correctly identifying entity boundaries when an adjective precedes a noun. The model may recognize part of the entity but fail to capture the full entity, or it may misclassify the type of the entity based on this common structure.

Complex Patterns

In addition to the simpler patterns, we also observed more complex structures, such as 'NOUN PUNCT ADJ' and 'NOUN NOUN'. These patterns are less frequent but still significant, indicating scenarios where the model needs to handle more intricate syntactic structures.

CHAPTER 4

Conclusion

This thesis investigated the robustness and generalizability of two Named Entity Recognition (NER) models, BioBERT and KeBioLM, across different biomedical datasets. We aimed to understand how well these models perform when subjected to cross-dataset evaluations, which is critical for their application in real-world biomedical text mining tasks. Utilizing benchmark datasets BC5CDR and NCBI, we investigated the precision, recall, and F1 scores of each model. While both models demonstrate high accuracy within a single dataset, their performance drops significantly when applied to a different dataset. This drop highlights the challenges in generalizing across different biomedical corpora. We also filtered the sentences who were not part of the training dataset while testing. This showed great improvement in accuracy indicating that the models cannot generalize very well with unseen data.

The results of this thesis point to a number of possibilities for research. Investigating complex pre-training techniques, such as domain-adaptive pre-training, to enhance model generalization across various datasets is one potential path. Developing strategies that allow NER models to dynamically adjust to new domains while reducing the requirement for intensive retraining is another important subject.

List of Tables

Bibliography

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[DLL14] Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, 2014.

[K+21] Mihir P Khambete et al. Quantification of bert diagnosis generalizability across medical specialties using semantic dataset distance, May 2021.

[KK22] Hyunjae Kim and Jaewoo Kang. How do your biomedical named entity recognition models generalize to novel entities? *Ieee Access*, 10:31513–31523, 2022.

[LSJ+16] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068, 05 2016.

[LXY+21] Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. Crossner: Evaluating cross-domain named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13452–13460, May 2021.

[LYK+19] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019.

[NS07] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[YLT+21] Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. Improving biomedical pretrained language models with knowledge, 2021.