Biomedical Entity Recognition Using Distant Supervision and Transformer-Based Models

Habibatou Allah Amizmiz
Department of Computer Science (FSSM)
Faculty of Sciences Semlalia, Cadi Ayyad University
Marrakesh, Morocco
h.amizmiz1698@uca.ac.ma

Hafsa Ikram
Department of Computer Science (FSSM)
Faculty of Sciences Semlalia, Cadi Ayyad University
Marrakesh, Morocco
h.ikram2758@uca.ac.ma

JIHAD Zahir
Department of Computer Science (FSSM)
Faculty of Sciences Semlalia, Cadi Ayyad University
Marrakesh, Morocco
j.zahir@uca.ac.ma

***Abstract*—Extracting biomedical information from scientific texts remains a major challenge due to the complexity and growing volume of available data. This work proposes an automated approach for biomedical Named Entity Recognition (NER), targeting genes, proteins, symptoms, and diseases. Relying on distant supervision, pre-trained language models are fine-tuned using an automatically annotated corpus. The study demonstrates the effectiveness of this approach in rapidly generating high-quality annotated data and improving model performance on specialized medical texts.**

I. Introduction

The biomedical field is experiencing an exponential growth in the volume of scientific publications, offering a wealth of valuable but difficult-to-manually-process information [1]. In this context, Named Entity Recognition (NER), which aims to automatically identify and classify entities such as diseases, genes, proteins, or clinical symptoms, emerges as a key tool for knowledge extraction, clinical decision support, and drug discovery [2].

Traditional NER approaches rely on supervised learning, requiring manually annotated corpora [3]. However, annotation in the biomedical domain is particularly costly, time-consuming, and demands specialized expertise [4]. This constitutes a major barrier to creating large-scale training datasets for each sub-domain.

To address this constraint, we propose a comprehensive biomedical NER method based on distant supervision [5], combining external resources (entity lists) and linguistic rules to automatically annotate a corpus of scientific abstracts extracted from PubMed [6]. The focus is on four types of biomedical entities: genes, proteins, diseases, and symptoms.

II. Work Flow

The workflow for biomedical Named Entity Recognition is structured as follows:

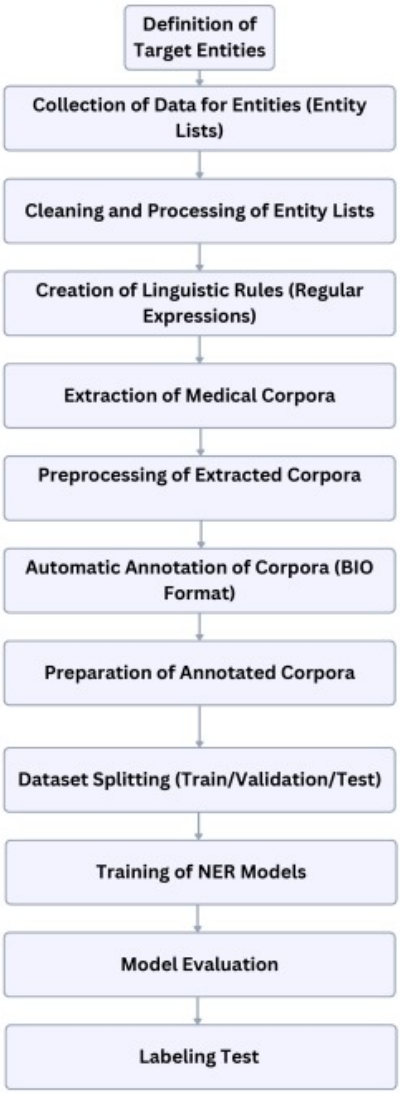

Fig. 1. Workflow for biomedical NER.

III. Data Description

A. Data Source

The entities used in this study were obtained from reliable biomedical reference resources:

- **Genes**: Human Phenotype Ontology (HPO) [7]
- **Proteins**: HUGO Gene Nomenclature Committee (HGNC) [8]
- **Symptoms and Diseases**: SNOMED CT ontology accessed via BioPortal [9]

These databases are widely recognized in the biomedical field for their reliability and standardized ontological structure [10].

B. Type and Format

TABLE I
Type and format of entity lists

Category	Data Type	Format	Encoding	Content
Genes	List of genetic symbols	Plain text (.txt)	UTF-8	1 symbol per line (e.g., BRCA1)
Proteins	Protein names	Tab-separated (.tsv)	UTF-8	Extraction of the `name` column
Symptoms	Clinical terms	Plain text (.txt)	UTF-8	1 symptomatic term per line
Diseases	Clinical terms	Plain text (.txt)	UTF-8	1 disease name per line

C. Data Volume

TABLE II
Data volume by category

Category	Number of Unique Entities
Genes	18,204
Proteins	19,268
Symptoms	248
Diseases	248

D. Collection and Scraping

Entity collection was performed using both manual and automated procedures depending on the data type.

Genes were retrieved manually from the HPO portal, from files containing genetic annotations associated with phenotypes and diseases. These files were merged locally to obtain a unified list.

Proteins were extracted from the official HGNC website. The entire tabulated file was downloaded, and only the field corresponding to the full protein name (`name`) was retained to form the final dataset.

Symptoms and **diseases** were collected automatically by querying the BioPortal REST API [11]. A Python script was used to send targeted requests to the SNOMED CT ontology to retrieve relevant clinical concepts. The JSON responses were parsed and saved locally as text lists.

E. Preprocessing

A standardized preprocessing was applied to each entity list to ensure consistency and usability for distant supervision.

For **genes**, the case was preserved, as genetic symbols are case-sensitive. For **proteins**, **symptoms**, and **diseases**, all terms were converted to lowercase to facilitate detection in texts.

Each file underwent a series of operations: removal of duplicates, elimination of empty lines, cleaning of non-alphanumeric special characters, and uniform UTF-8 encoding. The resulting lists are compact, homogeneous, and directly usable in the rule-based and regex-driven annotation process.

This processing ensured minimal quality and the necessary standardization for reliable and reproducible automatic annotation.

IV. Data Annotation

Automatic corpus annotation is a central component of our biomedical Named Entity Recognition pipeline. To reduce reliance on costly and expertise-intensive manual annotation, we adopted a distant supervision approach [12], combining existing lexical resources and linguistic rules based on frequent patterns in medical texts.

A. Source of Medical Articles

The annotated text documents were automatically collected from the PubMed scientific database by querying the Entrez API provided by the NCBI [13]. To target the four entity types considered in this study—genes, proteins, symptoms, and diseases—four specific queries were defined: “human gene”, “human protein”, “disease symptoms”, and “human diseases”. For each query, up to 80 abstracts were retrieved in raw format. Each document was associated with a unique identifier (PMID) and labeled according to its source category. This process created a diverse and directly usable corpus for automatic annotation.

B. Preprocessing of Abstracts

Before annotation, the texts underwent a standardized preprocessing phase. Each abstract was cleaned by converting to lowercase, removing punctuation, non-alphanumeric characters, and stop words. Orthographic normalization was also applied to standardize representations. Sentence and token segmentation prepared the data for word-by-word labeling in the BIO format [14].

C. Distant Supervision Annotation Strategy

Our approach relies on a hybrid strategy combining two complementary mechanisms. On one hand, pre-cleaned external entity lists (genes, proteins, symptoms, and diseases) were used to identify exact matches in the text. On the other hand, linguistic rules based on regular expressions were developed to detect lexical variants not covered by the lists. This combination ensures better coverage while limiting annotation errors.

D. Annotation Process Steps

- **Text Preprocessing**: Conversion to lowercase, removal of punctuation and stop words, and sentence segmentation.
- **Tokenization**: Splitting sentences into individual tokens compatible with the model’s tokenizer.
- **Entity Detection**: Identifying occurrences through exact matching with entity lists, supplemented by specific regex rules.

- **BIO Labeling**: Annotating each token with `B-`, `I-`, or `O` labels based on its position in the recognized entity.
- **Alignment and Export**: Saving annotated data in standard CoNLL/BIO format, with a (token, label) pair per line.

V. Models

In this study, we evaluated two pre-trained language models specialized in the biomedical domain: **BioBERT** [15] and **PubMedBERT** [16]. Both models are based on the *BERT base* architecture (Bidirectional Encoder Representations from Transformers) [17], widely used for Named Entity Recognition (NER) tasks.

A. Data Preparation

The texts were automatically annotated according to the BIO (*Begin, Inside, Outside*) scheme, identifying entities of type **gene**, **protein**, **disease**, and **symptom**. Each sentence was:

- segmented, cleaned, and tokenized using the model's associated tokenizer;
- aligned with corresponding BIO labels;
- encoded with `-100` to ignore non-annotated tokens (`O`) in the loss function.

The final corpus was divided into three subsets:

- Training (`train`)
- Validation (`val`)
- Test (`test`)

B. BioBERT

- **Model Used**: `dmis-lab/biobert-base-cased-v1.1`
- **Pre-training**: Wikipedia, PubMed abstracts, PMC full texts
- **Type**: *Cased* (case-sensitive, relevant for genetic symbols)
- **Tokenizer**: WordPiece
- **Architecture**: 12 layers, 768 hidden dimensions, 12 attention heads
- **Number of Parameters**: ˜110M

C. PubMedBERT

- **Model Used**: `microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract`
- **Pre-training**: Exclusively on PubMed abstracts
- **Type**: *Uncased* (case-insensitive)
- **Tokenizer**: WordPiece with PubMed-specialized vocabulary
- **Architecture**: 12 layers, 768 hidden dimensions, 12 attention heads
- **Number of Parameters**: ˜110M

D. Training Parameters

Both models were fine-tuned on our annotated corpus using the `Trainer` API from `HuggingFace Transformers` [18]. The hyperparameters used are listed in Table III.

Performance was measured using standard NER metrics: **precision**, **recall**, and **F1-score**, calculated with the `seqeval` library [19].

TABLE III
HYPERPARAMETERS USED DURING FINE-TUNING

Parameter	Value
Optimizer	AdamW
Learning rate	2×10^{-5}
Batch size	8
Number of epochs	4
Evaluation	Every 500 steps
Regularization (weight decay)	0.01
Maximum sequence length	512
Ignore "O" labels	`-100`

VI. Evaluation, Results, and Discussion

A. Evaluation Protocol

Both models, PubMedBERT and BioBERT, were evaluated on two datasets: a validation set and a test set. The metrics used were:

- **Accuracy**: Proportion of correctly labeled tokens.
- **F1-score**: Harmonic mean of precision and recall, calculated per class and globally.
- **Micro, macro, and weighted averages**: To assess balance across classes.

Evaluation was performed using the `seqeval` library, considering only annotated entities (BIO).

B. Validation Set Results

TABLE IV
VALIDATION SCORES BY MODEL AND ENTITY

Entity	Model	Precision	Recall	F1-score
DISEASE	PubMedBERT	0.99	0.98	0.98
	BioBERT	0.98	0.99	0.99
GENE	PubMedBERT	0.87	0.97	0.92
	BioBERT	0.84	0.94	0.89
PROTEIN	PubMedBERT	0.92	0.85	0.88
	BioBERT	0.94	0.87	0.90
SYMPTOM	PubMedBERT	1.00	1.00	1.00
	BioBERT	1.00	0.60	0.75
Global F1	PubMedBERT	—	—	**0.9407**
	BioBERT	—	—	**0.9380**

C. Test Set Results

TABLE V
TEST SCORES BY MODEL

Entity	Model	Precision	Recall	F1-score
DISEASE	PubMedBERT	0.98	1.00	0.99
	BioBERT	0.97	1.00	0.98
GENE	PubMedBERT	0.83	0.95	0.88
	BioBERT	0.96	0.96	0.96
PROTEIN	PubMedBERT	0.96	0.84	0.90
	BioBERT	0.97	0.94	0.95
SYMPTOM	PubMedBERT	1.00	1.00	1.00
	BioBERT	1.00	0.92	0.96
Global F1	PubMedBERT	—	—	**0.9395**
	BioBERT	—	—	**0.9689**

D. Ablation Study on BioBERT

To analyze the impact of dataset size and model configurations on BioBERT's performance, two ablation studies were conducted.

1) Ablation on Dataset Size: The first study examined the effect of different training, validation, and test set splits, maintaining a total of 320 data lines. Five configurations were tested: 80%-10%-10% (256/32/32 lines), 70%-15%-15% (224/48/48 lines), 90%-5%-5% (288/16/16 lines), 60%-20%-20% (192/64/64 lines), and 75%-12.5%-12.5% (240/40/40 lines). The results are summarized in Table VI. The 90%-5%-5% configuration achieves a perfect F1-score (1.00) on the validation set, likely due to the small validation set size, but its test set performance is slightly lower (0.9417). The 60%-20%-20% configuration shows the best test set F1-score (0.9710), suggesting good generalization despite a reduced training set.

TABLE VI
ABLATION STUDY RESULTS ON DATA SPLITS FOR BIOBERT

Configuration	Training	Validation (Acc/F1)	Test (Acc/F1)
80%-10%-10%	256 lines	0.9380 / 0.9380	0.9689 / 0.9689
70%-15%-15%	224 lines	0.9549 / 0.9549	0.9563 / 0.9563
90%-5%-5%	288 lines	1.0000 / 1.0000	0.9417 / 0.9417
60%-20%-20%	192 lines	0.9110 / 0.9110	0.9710 / 0.9710
75%-12.5%-12.5%	240 lines	0.9414 / 0.9414	0.9706 / 0.9706

2) Ablation on Model Configurations: A second ablation study was conducted on the 60%-20%-20% configuration (192 training lines, 64 validation lines, 64 test lines) to analyze the impact of architectural modifications on BioBERT. Five variants were tested:

- **Baseline**: Standard BioBERT model without modifications.
- **Frozen Layers**: Freezing the first six layers of the model to limit weight adjustments.
- **High Dropout**: Increasing the dropout rate to 0.3 for stronger regularization.
- **Simplified Head**: Replacing the standard classification head with a single linear layer.
- **Random Weights**: Initializing the model with random weights, without pre-training.

The results are presented in Table VII. The baseline configuration achieves an F1-score of 0.9092 on validation and 0.9789 on test, outperforming all modified variants. The frozen layers variant (0.8866 on validation, 0.9604 on test) and the high dropout variant (0.8761 on validation, 0.9578 on test) show a slight performance drop, suggesting that limiting layer adjustments or increasing regularization may reduce the model's ability to capture complex patterns. The simplified classification head (0.8743 on validation, 0.9551 on test) performs similarly to high dropout. The random weights variant yields the lowest performance (0.8080 on validation, 0.9129 on test), highlighting the critical importance of pre-training for biomedical entity recognition [20].

TABLE VII
ABLATION STUDY RESULTS ON MODEL CONFIGURATIONS FOR BIOBERT (60%-20%-20%)

Configuration	Validation (Acc/F1)	Test (Acc/F1)
Baseline	0.9092 / 0.9092	0.9789 / 0.9789
Frozen Layers	0.8866 / 0.8866	0.9604 / 0.9604
High Dropout (0.3)	0.8761 / 0.8761	0.9578 / 0.9578
Simplified Head	0.8743 / 0.8743	0.9551 / 0.9551
Random Weights	0.8080 / 0.8080	0.9129 / 0.9129

E. Project Limitations

Despite the promising performance of our approach, several limitations must be highlighted:

- **Quality of Distant Supervision Annotation**: The use of external entity lists and regex-based rules may introduce noise in annotations, particularly for ambiguous or context-dependent entities (e.g., terms like "pain" that can be a symptom or a generic word) [21]. This can affect model precision, especially for less frequent entities like symptoms (see Table IV).
- **Limited Corpus Size**: With only 320 annotated data lines (Table VI), the corpus remains relatively small compared to the needs of large-scale language models. This limitation may restrict the models' ability to generalize to diverse biomedical texts, particularly for specific subdomains [22].
- **Entity Coverage**: The entity lists used (Table II) cover a limited number of symptoms and diseases (248 each) compared to genes (18,204) and proteins (19,268). This may lead to imbalances in entity detection, as observed in the variable performance for symptoms (Table V).
- **Generalization to Non-Standardized Texts**: The models were trained on PubMed scientific abstracts, which are relatively structured. Their performance may decrease on less formal texts, such as medical records or patient forums, due to lexical and stylistic variations [23].
- **Dependence on Pre-training**: The ablation study on model configurations (Table VII) shows a strong dependence on pre-training, as evidenced by the performance drop with random weights. This limits the model's flexibility for emerging biomedical domains not covered by pre-training corpora.

These limitations suggest directions for future work, including improving annotation quality, increasing the size and diversity of the corpus, and adapting models to non-standardized biomedical texts.

F. Discussion

Analysis of the results reveals that:

- **BioBERT** slightly outperforms PubMedBERT on the test set with a global F1 of **0.9689** compared to **0.9395**, particularly on *GENE* and *PROTEIN* entities.
- **PubMedBERT** shows consistent and high performance on the *SYMPTOM* class (F1 = 1.00) on both sets, likely due to the lexical clarity of symptom mentions.

- For the *DISEASE* class, both models achieve very high F1 scores (0.98), indicating that diseases are well-identified.
- However, BioBERT appears more robust on the test set than on validation, suggesting better generalization, likely due to its mixed pre-training (Wikipedia + biomedical) [15].

In summary, while PubMedBERT performs well on validation, BioBERT achieves the best final results on the test data.

VII. Conclusion

This study presents a comprehensive approach to biomedical Named Entity Recognition using distant supervision and transformer-based models. The combination of external knowledge bases and linguistic rules for automatic annotation proves effective in generating high-quality training data. BioBERT demonstrates superior performance with an F1-score of 0.9689 on the test set, outperforming PubMedBERT across most entity types. The ablation studies reveal the importance of pre-training and optimal dataset configurations. Future work should focus on expanding the corpus size, improving annotation quality, and enhancing generalization to diverse biomedical text types.

References

[1] Q. Chen, A. Peng, and J. Laurent, "Discourse-level coherence in the PubMed corpus," *PLOS ONE*, vol. 13, no. 2, p. e0193878, 2018.

[2] N. Kang, Z. Singh, A. Afzal, E. M. van Mulligen, and J. A. Kors, "Using rule-based natural language processing to improve disease normalization in biomedical text," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 876–881, 2013.

[3] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu, "BioCreative V CDR task corpus: a resource for chemical disease relation extraction," *Database*, vol. 2016, p. baw068, 2016.

[4] X. Wang, Y. Zhang, X. Ren, Y. Zhang, M. Zitnik, J. Shang, C. Langlotz, and J. Han, "Cross-type biomedical named entity recognition with deep multi-task learning," *Bioinformatics*, vol. 35, no. 10, pp. 1745–1752, 2018.

[5] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 2009, pp. 1003–1011.

[6] NCBI Resource Coordinators, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 52, no. D1, pp. D33–D43, 2024.

[7] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, "The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease," *The American Journal of Human Genetics*, vol. 83, no. 5, pp. 610–615, 2008.

[8] S. Povey, R. Lovering, E. Bruford, M. Wright, M. Lush, and H. Wain, "The HUGO Gene Nomenclature Committee (HGNC)," *Human Genetics*, vol. 109, no. 6, pp. 678–680, 2001.

[9] N. F. Whetzel, N. H. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen, "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications," *Nucleic Acids Research*, vol. 39, no. suppl_2, pp. W541–W545, 2011.

[10] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[11] M. Salvadores, P. R. Alexander, M. A. Musen, and N. F. Noy, "BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF," *Semantic Web*, vol. 4, no. 3, pp. 277–284, 2013.

[12] M. Craven and J. Kumlien, "Constructing biological knowledge bases by extracting information from text sources," in *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, 1999, pp. 77–86.

[13] E. W. Sayers, M. Cavanaugh, K. Clark, K. J. Ostell, K. D. Pruitt, and I. Karsch-Mizrachi, "GenBank," *Nucleic Acids Research*, vol. 52, no. D1, pp. D62–D69, 2024.

[14] L. A. Ramshaw and M. P. Marcus, "Text chunking using transformation-based learning," in *Third Workshop on Very Large Corpora*, 1995, pp. 82–94.

[15] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[16] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–23, 2021.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[18] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[19] H. Nakayama, "seqeval: A Python framework for sequence labeling evaluation," Software available from `https://github.com/chakki-works/seqeval`, 2018.

[20] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[21] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld, "Knowledge-based weak supervision for information extraction of overlapping relations," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 541–550.

[22] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *China National Conference on Chinese Computational Linguistics*, 2019, pp. 194–206.

[23] S. Henry, K. Buchan, M. Filannino, A. Stubbs, and O. Uzuner, "2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records," *Journal of the American Medical Informatics Association*, vol. 27, no. 1, pp. 3–12, 2020.