Biomedical Named Entity Recognition and Normalization Tools

RWTH Aachen University
Hanbin Chen

RWTH AACHEN UNIVERSITY

Overview

- **1. Background**
  - BTM, NER, NEN
- **2. Challenges**
- **3. Models and Corpora**
- **4. Evaluation Method and Result**
- **5. Discussion**
- **6. Summary**

PAPER

HunFlair2 in a cross-corpus evaluation of biomedical named entity recognition and normalization tools

Mario Sänger[1,*,†] Samuele Garda,[1,†] Xing David Wang,[1,†] Leon Weber-Genzel,[2] Pia Droop,[1] Benedikt Fuchs,[3] Alan Akbik[1] and Ulf Leser[1,*]

[1]Department of Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany, [2]Center for Information and Language Processing (CIS), Ludwig Maximilian University Munich, Geschwister-Scholl-Platz 1, 80539 München, Germany and [3], Research Industrial Systems Engineering (RISE) Forschungs-, Entwicklungs- und Großprojektberatung GmbH, Concorde Business Park F, 2320 Schwechat, Austria

*Corresponding authors: saengema@informatik.hu-berlin.de and leser@informatik.hu-berlin.de

†Authors contributed equally.

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

1. Background

- **1. Background**
  - BTM, NER, NEN
- **2. Challenges**
- **3. Models and Corpora**
- **4. Evaluation Method and Result**
- **5. Discussion**
- **6. Summary**

1. Background

- **Biomedical Text Mining (BTM)**
  - Extracts information from bio-literature

- **Key Processes**
  - NER (Named Entity Recognition)
    - Diseases, drugs, genes
  - NEN (Named Entity Normalization)
    - Link entities to standard Knowledge base(KS) / dictionary
      - NCBI (National Center for Biotechnology Information)
      - CTD (Comparative Toxicogenomics Database)

- **Challenges**
  - Ambiguity, complex terminology
  - AI advancements, analytical integration

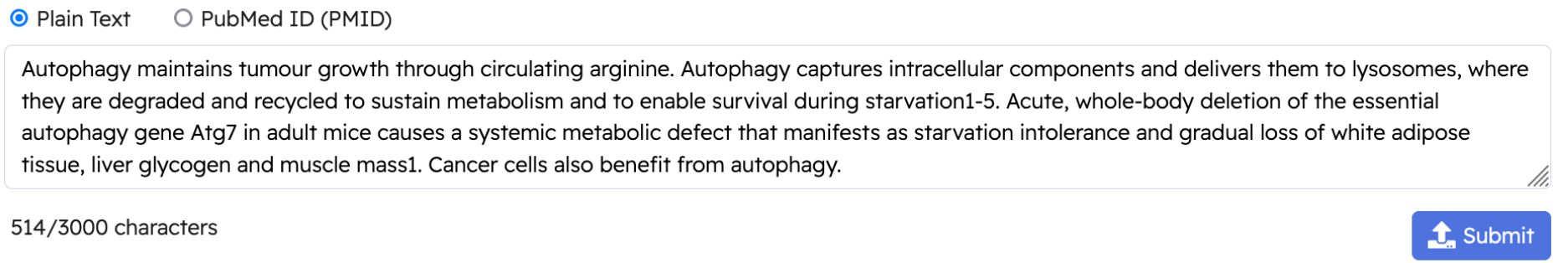
NIH National Library of Medicine
National Center for Biotechnology Information

ctd *Illuminating how chemicals affect human health.*
Comparative Toxicogenomics Database

https://www.ncbi.nlm.nih.gov
https://ctdbase.org

1. Background

- BERN2 for example



2. Challenges

- **1. Background**
  - BTM, NER, NEN
- **2. Challenges**
- **3. Models and Corpora**
- **4. Evaluation Method and Result**
- **5. Discussion**
- **6. Summary**

2. Challenge

- **Non-consecutive / Overlapping**

“[...] is causing breast and ovarian cancer [...]”

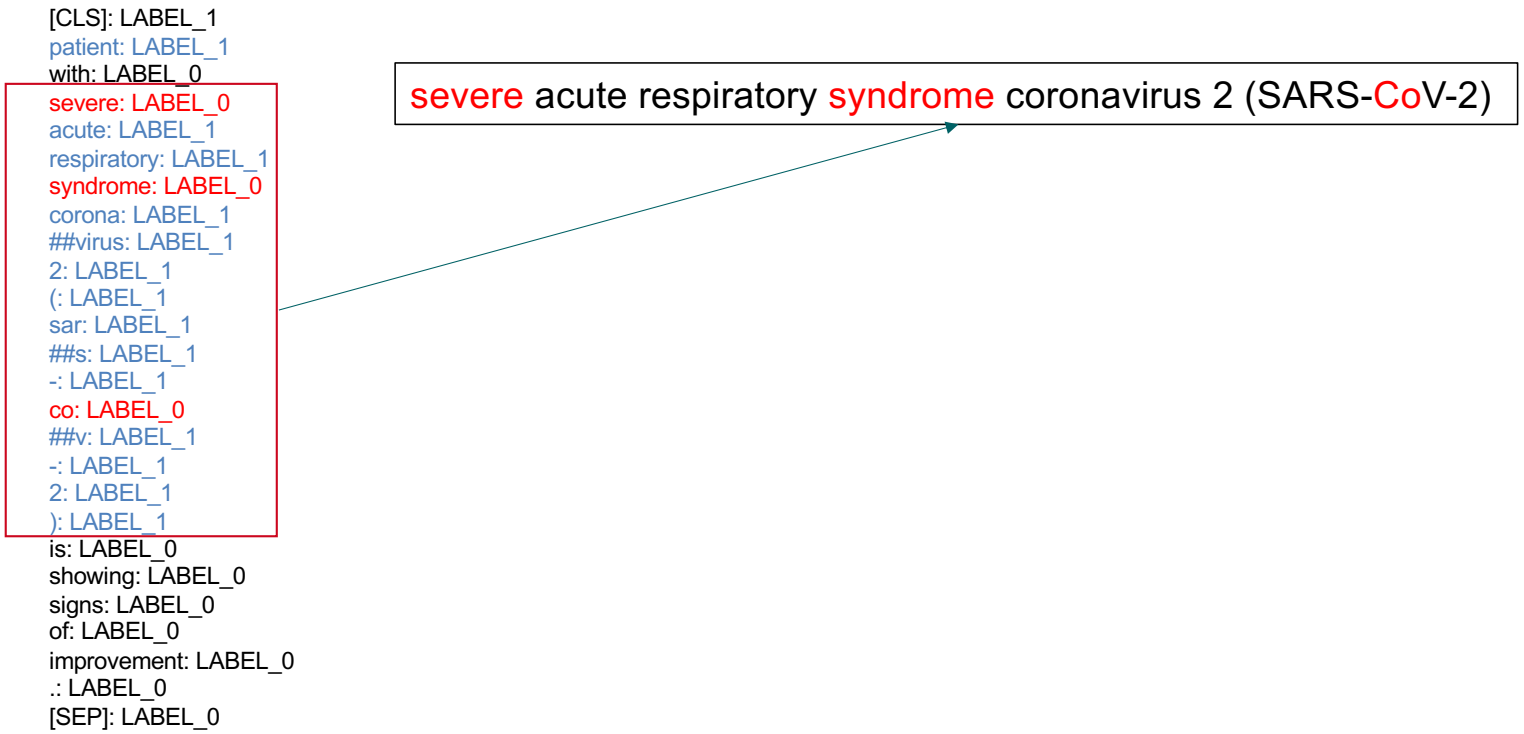
  - “breast” and “ovarian cancer ”
  - “breast and ovarian cancer ”
  - “breast cancer ” and “ovarian cancer

- **Synonyms**
  - **Diabetes**: Most commonly used term.
  - **Diabetes mellitus**: The formal medical term
  - **DM**: Abbreviation for "Diabetes Mellitus"
  - **Hyperglycemia**:
    - Sometimes used in the context of describing prediabetes or complications, although it primarily describes a symptom.
  - **Type 1 Diabetes** and **Type 2 Diabetes**:

2. Challenge

Example of BlueBERT handling NER:

"Patient with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is showing signs of improvement."



2. Challenge

- **Data Quality and Availability Status:**
  - BC2GM (2007): BioCreative II Gene Mention
  - BC4CHEMD (2013)
  - Linnaeus Dataset (2010)
  - CRAFT (2012)
  - BioNLP13 CG (2013)

- **Data Imbalance in Biomedical Research**
  - rare diseases with limited descriptions

- **Explainability**
  - The "black box" nature of LLM technology

- **Existing Benchmark Limitations**
  - Focus: Only Recognition or Normalization
  - Lacks: End-to-End NER and NEN Results

- **Technological Updates**
  - Ignored: Latest Transformer-Based Models

3. Models and Corpora

- **1. Background**
  - BTM, NER, NEN
- **2. Challenges**
- **3. Models and Corpora**
- **4. Evaluation Method and Result**
- **5. Discussion**
- **6. Summary**

PAPER

HunFlair2 in a cross-corpus evaluation of biomedical named entity recognition and normalization tools

Mario Sänger,[1,*,†] Samuele Garda,[1,†] Xing David Wang,[1,†] Leon Weber-Genzel,[2] Pia Droop,[1] Benedikt Fuchs,[3] Alan Akbik[1] and Ulf Leser[1,*]

[1]Department of Computer Science, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany, [2]Center for Information and Language Processing (CIS), Ludwig Maximilian University Munich, Geschwister-Scholl-Platz 1, 80539 München, Germany and [3], Research Industrial Systems Engineering (RISE) Forschungs-, Entwicklungs- und Großprojektberatung GmbH, Concorde Business Park F, 2320 Schwechat, Austria

*Corresponding authors: saengema@informatik.hu-berlin.de and leser@informatik.hu-berlin.de

†Authors contributed equally.

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

3. Models

- **Model Selection Criteria**
  - C1: Supports both NER and NEN
  - C2: Utilizes machine-learning-based models for NER
    - Machine-learning NER as state-of-the-art
  - C3: Extracts genes, diseases, chemicals, species
    - Important for downstream applications
  - C4: No additional licenses required (e.g., commercial, UMLS)
    - Usability in research pipelines without licensing constraints

- **Qualified Tools**
  - BERN2
  - PubTator
  - SciSpacy
  - bent
  - HunFlair2

3. Models

Table 1. Overview of the tools selected for our evaluation. We distinguish rule-based ("RB"), machine learning-based ("ML") and neural-network based ("NN") approaches for NER and NEN. Moreover, for each tool we illustrate the support of the following entity types: genes (Ge), species (Sp), disease (Di), chemical (Ch), cell line (Cl) and variant (Va). For each entity type we illustrate whether the tool supports NER and NEN of the type by marking the column with ✓, if only NER is supported we use (✓). Last update highlights the last update of the code repository of the respective tool. Citations counts are taken from Google Scholar on 01/10/2024.

Tool	API	Ge	Sp	Di	Ch	Cl	Va	NER	NEN	Pub. Year	Last Update	Citations
PubTator Central [116]	REST/ Tools	✓	✓	✓	✓	✓	✓	ML / NN	RB	2019	-	315
BERN2 [78]	REST/ Python	✓	✓	✓	✓	✓	✓	NN	RB / NN	2022	11/2023	46
SciSpacy [79]	Python	(✓)	✓	✓	✓	✓	(✓)	NN	RB	2019	10/2023	635
bent [88, 89]	Python	✓	✓	✓	✓	✓	(✓)	NN	RB	2020	12/2023	13
HunFlair2 [114]	Python	✓	✓	✓	✓	✓		NN	RB / NN	2021	01/2024	83

Sänger, Mario, et al. "HunFlair2 in a cross-corpus evaluation of named entity recognition and normalization tools." *arXiv preprint arXiv:2402.12372* (2024).

3.1 BERN2

NER (Named Entity Recognition)
- Transformer-Based: **RoBERTa**
  - Multi-Task Training

NEN (Named Entity Normalization)
- Hybrid System:
  - Rule-Based + Neural-Based
- Neural-Based Model: BioSyn

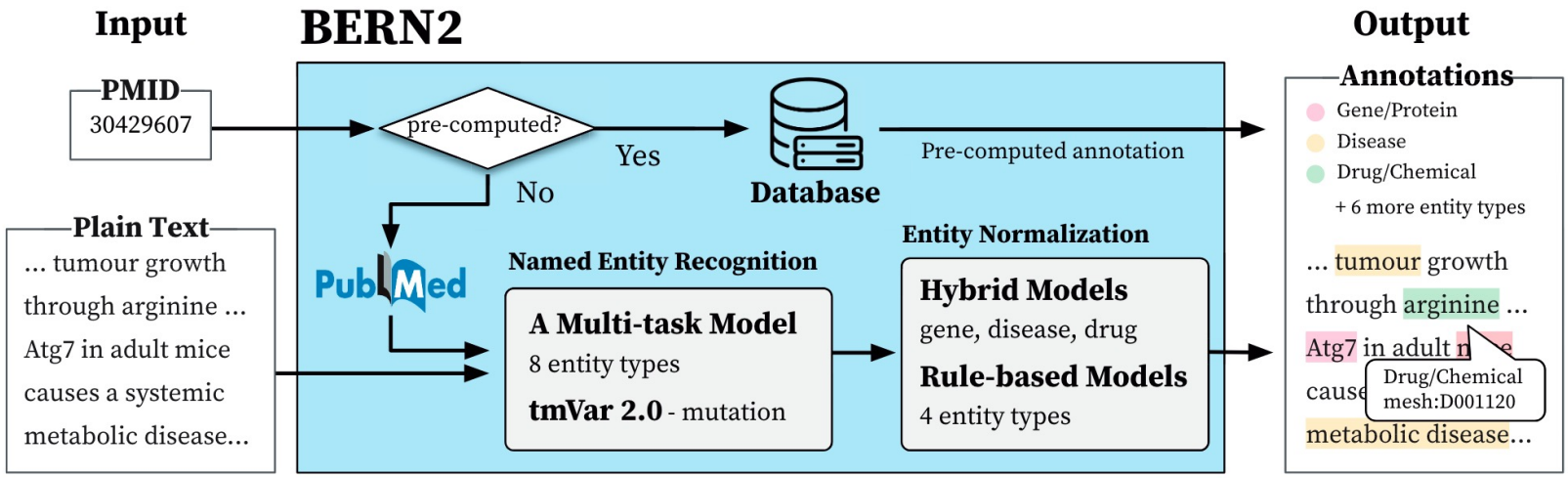


Fig. 1. An overview of BERN2. Given plain text or a PubMed ID (PMID), BERN2 recognizes nine biomedical entity types and normalizes each concept

Sung, Mujeen, et al. "BERN2: an advanced neural biomedical named entity recognition and normalization tool." *Bioinformatics* 38.20 (2022): 4837-4839.

3.1 BERN2

Entity Type	NER Model	NER Training Corpus	NEN Model	NEN Training Corpus
Genes	RoBERTa	BC2GM	GNormPlus, BioSyn(NN)	BC2GN
Diseases		NCBI Disease	sieve-based approach, BioSyn(NN)	BC5CDR, NCBI Disease
Chemicals		BC4CHEMD	tmChem4, BioSyn(NN)	BC5CDR
Species		Linnaeus	dictionary lookup	Not specified for species

3.2 PubTator

Entity Type	NER Model	NER Training Corpus	NEN Model	NEN Training Corpus
Genes	BlueBERT	GNormPlus, NLM-Gene	TF-IDF frequencies	Not specifically stated
Species	SR4GN (Rule-based system)	Not specifically stated	SR4GN (Rule-based system)	Not specifically stated
Chemicals	BlueBERT	BC5CDR, NLM-Chem	Multi-terminology candidate resolution (MTCR)	Not specifically stated
Disease	TaggerOne	NCBI Disease, BC5CDR corpora	TaggerOne	NCBI Disease, BC5CDR corpora

SR4GN (Species Recognition for Gene Normalization)

3.3 SciSpacy

Entity Type	NER Model	NER Training Corpus	NEN Model
Genes	Stack LSTMs(NN)	CRAFT, BioNLP13 CG	string-matching approach based on characters 3-grams
Diseases	Stack LSTMs(NN)	BC5CDR, BioNLP13 CG	
Chemicals	Stack LSTMs(NN)	BC5CDR, CRAFT, BioNLP13 CG	
Species	Stack LSTMs(NN)	CRAFT, BioNLP13 CG	

Stack LSTMs (Stacked Long Short-Term Memory networks)

3.4 bent

Entity Type	NER Model	NER Training Corpus	NEN Model
Genes	PubMedBERT	BC2GM, CRAFT	PageRank
Diseases	PubMedBERT	BC5CDR, NCBI-disease	
Chemicals	PubMedBERT	BC5CDR, NLMChem	
Species	PubMedBERT	Linnaeus, CRAFT	

3.5 HunFlair2

- **NER (Named Entity Recognition)**
  - Entity Extraction:
    - BioLink-BERT
    - Joint Extraction, the end-to-end entity extraction
- **Indicate the entity types to extract**
  - Examples:
    - [Tag genes] <input-example>
    - [Tag diseases] <input-example>
    - [Tag chemicals, diseases, genes, species] <input-example>
  - Output Labels: IOB Scheme (B-<entity type>, I-<entity type>)

- **NEN (Named Entity Normalization)**
  - Models Employed:
    - BioSyn
    - SapBERT

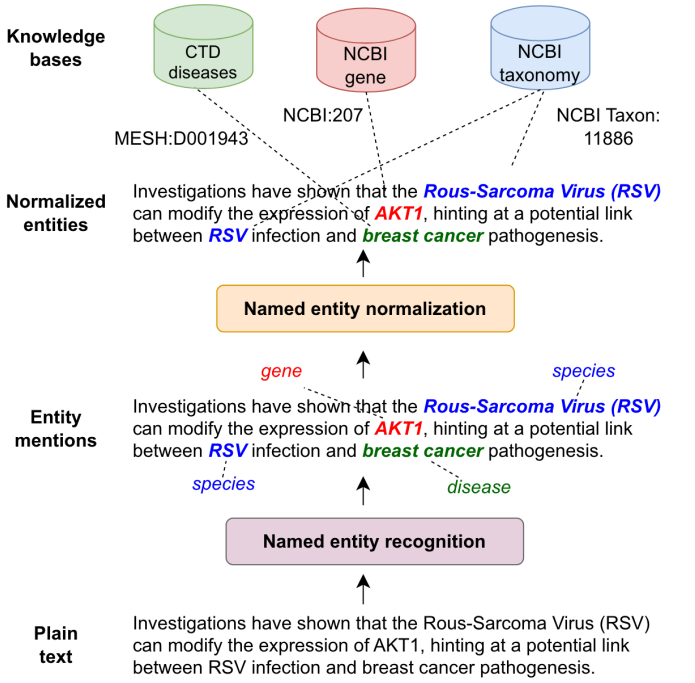


Fig. 1. Illustration of the named entity extraction process. First entity mentions in plain will be identified using named entity recognition (NER) tools. Afterwards named entity normalization (NEN) approaches map the found mentions to standard identifiers in a knowledge base.

Sänger, Mario, et al. "HunFlair2 in a cross-corpus evaluation of named entity recognition and normalization tools." *arXiv preprint arXiv:2402.12372* (2024).

3.5 HunFlair2

Entity Type	NER Model	NER Training Corpus	NEN Model	NEN Training Corpus
Genes	BioLink-BERT	BioRED, NLM Gene, GNormPlus	BioSyn	BC2GN, NCBI Gene (human subset)
Diseases		BioRED, NCBI Disease, SCAI Disease		Not explicitly mentioned for diseases
Chemicals		BioRED, NLM Chem, SCAI Chemical		Not explicitly mentioned for chemicals
Species		BioRED, Linneaus, S800	SapBERT	UMLS

3. Models and Corpora

Model Name	NER Technique Details	NEN Technique Details
BERN2	RoBERTa	Rule-based, Neural Networks
bent	PubMedBERT	PageRank algorithm
PubTator	BlueBERT, rule-based	TF-IDF frequencies mixed methods
SciSpacy	Stack LSTMs	Character 3-grams string matching
HunFlair2	BioLink-BERT	Neural Networks, SapBERT

3. Models and Corpora

Models	Genes	Chemicals	Diseases	Species
BERN2	BC2GM	BC4CHEMD	NCBI Disease	Linnaeus
bent	BC2GM, CRAFT	BC5CDR, NLM-Chem	BC5CDR, NCBI-disease	Linnaeus, CRAFT
PubTator	GnormPlus, NLM-Gene	BC5CDR, NLM-Chem	NCBI-Disease, BC5CDR	-
SciSpacy	CRAFT	BC5CDR, BioNLP13 CG	BC5CDR	CRAFT, BioNLP13 CG
HunFlair2	NLM Gene, GNormPlus	NLM Chem, SCAI Chemical	NCBI Disease, SCAI disease	Linnaeus, S800

4. Evaluation Method and Result

- **1. Background**
  - BTM, NER, NEN
- **2. Challenges**
- **3. Models and Corpora**
- **4. Evaluation Method and Result**
- **5. Discussion**
- **6. Summary**

4.1 Evaluation

- **End-to-End Approach**: Direct entity and relation identification for efficiency
  - Normal benchmarks have limitations too, either recognizing entities or normalizing
  - End-to-End: start and end offset of the mention boundary and KB identifier. triplets (start, end, KBID)

- **Data Selection Criteria:**
  - a: Corpora unused in tool training (training/development split)
  - b: Corpora with NER and NEN annotations
  - c: Entity types normalized to universally supported KBs

- **Knowledge Bases (KBs) Selected:**
  - Genes: NCBI Gene
  - Diseases: CTD Diseases
  - Chemicals: CTD Chemicals
  - Species: NCBI Taxonomy

- **Corpora Selected for Benchmark:**
  - BioID
  - MedMentions
  - tmVar (v3)

4.2 Result

Tool	In-corpus	Cross-corpus
BERN2		
Chemical	96.60† *(BC5CDR)*	41.68 *(MedMentions)*
Disease	93.90† *(BC5CDR)*	47.31 *(MedMentions)*
Gene	95.90† *(BC2GM)*	43.81 *(tmVar v3)*
PubTator		
Chemical	77.20 *(NLM-Chem)*	31.26 *(MedMentions)*
Disease	80.70 *(NCBI-Disease)*	40.76 *(MedMentions)*
Gene	72.70 *(NLM-Gene)*	85.92 *(tmVar v3)*

Sänger, Mario, et al. "HunFlair2 in a cross-corpus evaluation of named entity recognition and normalization tools." *arXiv preprint arXiv:2402.12372* (2024).

4.2 Result

	BERN2	HunFlair2	PubTator	SciSpacy	bent
Chemical MedMentions	41.79 (33.42†)	**51.17**	31.28	34.95	40.90
Disease MedMentions	47.33	**57.57**	41.11	40.78	45.94
Gene tmVar (v3)	43.96	76.75	**86.02**	-	0.54
Species BioID	14.35	49.66	**58.90**	37.14	10.35
Avg	36.86 (34.72†)	**58.79**	54.33	37.61	24.43

Sänger, Mario, et al. "HunFlair2 in a cross-corpus evaluation of named entity recognition and normalization tools." *arXiv preprint arXiv:2402.12372* (2024).

5. Discussion

- **1. Background**
  - BTM, NER, NEN
- **2. Challenges**
- **3. Models and Corpora**
- **4. Evaluation Method and Result**
- **5. Discussion**
- **6. Summary**

5. Discussion

- **In-Corpus Evaluations:**
  - Consistency, high scores, limited real-world applicability.
- **Cross-Corpus Evaluations:**
  - Unfamiliar datasets, realistic generalization assessment.
  - Lower performance, generalization challenges.
- **Annotation Consistency:**
  - Varying guidelines and definitions, evaluation impact.
- **Evaluation Settings:**
  - excluding non-consecutive entities,
  - Method choices, potential tool bias.

5. Discussion

- **Multi-task LLM in BERN2**
  - Reduce Parameters
  - Enhanced Generalization
  - Increased Efficiency

- **Simplified LLM, DistilBERT**
  - Reduced Parameters
  - Faster Training
  - Performance Retention
  - Resource Optimization
  - Scalability
  - Biomedical Adaptability

6. Summary

- **1. Background**
  - BTM, NER, NEN
- **2. Challenges**
- **3. Models and Corpora**
- **4. Evaluation Method and Result**
- **5. Discussion**
- **6. Summary**

5. Summary

Study Focus:
- Biomedical NER and NEN tool evaluation
- Cross-corpus performance analysis

Tools Evaluated:
- HunFlair2, BERN2, bent, PubTator, SciSpacy

Evaluation Metrics:
- Precision, Recall, F1 Score
- End-to-End Approach

Main Findings:
- High performance in training-corpus
- Notable decline in cross-corpus settings
- Best performers: HunFlair2, BERN2

Challenges Identified:
- Generalization across different corpora
- Performance degradation in new contexts

Future Directions:
- Development of adaptive machine learning models
- Creation of diverse and comprehensive datasets
- Collaborative research efforts
- Simplified LLM
- Multi-task LLM in BERN2
- KB Enhancement

Conclusion:
- Need for innovations in model generalization
- Enhancement of biomedical text mining tools

References

- Sänger, Mario, et al. "HunFlair2 in a cross-corpus evaluation of named entity recognition and normalization tools." arXiv preprint arXiv:2402.12372 (2024).
- Sung, Mujeen, et al. "BERN2: an advanced neural biomedical named entity recognition and normalization tool." Bioinformatics 38.20 (2022): 4837-4839.
- Zhuang, Liu, et al. "A robustly optimized BERT pre-training approach with post-training." Proceedings of the 20th chinese national conference on computational linguistics. 2021.
- Kim, Donghyeon, et al. "A neural named entity recognition and multi-type normalization tool for biomedical text mining." IEEE Access 7 (2019): 73729-73740.
- Jain, Saahil, et al. "Radgraph: Extracting clinical entities and relations from radiology reports." arXiv preprint arXiv:2106.14463 (2021).
- Lewis, Patrick, et al. "Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art." Proceedings of the 3rd clinical natural language processing workshop. 2020.
- Wei, Chih-Hsuan, et al. "PubTator central: automated concept annotation for biomedical full text articles." Nucleic acids research 47.W1 (2019): W587-W593.

Thank you
for your attention !