A Comprehensive Evaluation of Biomedical Entity-centric Search

Elena Tutubalina
Insilico Medicine Hong Kong
elena@insilicomedicine.com

Zulfat Miftakhutdinov
Insilico Medicine Hong Kong
zulfat@insilicomedicine.com

Vladimir Muravlev
Insilico Medicine Hong Kong
v.muravlev@insilicomedicine.com

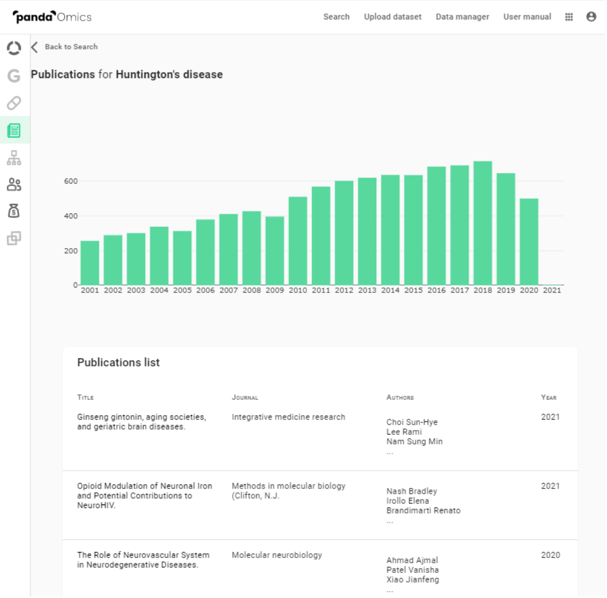
Anastasia Shneyderman
Insilico Medicine Hong Kong
a.shneyderman@insilicomedicine.com

Abstract

Biomedical information retrieval has often been studied as a task of detecting whether a system correctly detects entity spans and links these entities to concepts from a given terminology. Most academic research has focused on evaluation of named entity recognition (NER) and entity linking (EL) models which are key components to recognizing diseases and genes in PubMed abstracts. In this work, we perform a fine-grained evaluation intended to understand the efficiency of state-of-the-art BERT-based information extraction (IE) architecture as a biomedical search engine. We present a novel manually annotated dataset of abstracts for disease and gene search. The dataset contains 23K query-abstract pairs, where 152 queries are selected from logs of our target discovery platform and PubMed abstracts annotated with relevance judgments. Specifically, the query list also includes a subset of concepts with at least one ambiguous concept name. As a baseline, we use off-she-shelf Elasticsearch with BM25. Our experiments on NER, EL, and retrieval in a zero-shot setup show the neural IE architecture shows superior performance for both disease and gene concept queries.

1 Introduction

The amount of text data being produced is overwhelming, especially in biomedicine; PubMed[1] covers over 33 million articles from biomedical and life sciences journals and other texts, with about 1.5 million added each year. Meanwhile, many of these articles are about specific entities (e.g. proteins, diseases, chemicals), i.e., entity-centric. In general, entities are central to many search queries; e.g., (Guo et al., 2009) demonstrated that 71% of search queries contained named entities, while (Xiong et al., 2017) found that more than half of the traffic in the Allen Institute's scholar search engine is about research concepts.



Publications list

Title	Journal	Authors	Year
Ginseng gintonin, aging societies, and geriatric brain diseases.	Integrative medicine research	Choi Sun-Hye Lee Rami Nam Sung Min ...	2021
Opioid Modulation of Neuronal Iron and Potential Contributions to NeuroHIV.	Methods in molecular biology (Clifton, N.J.	Nash Bradley Irollo Elena Brandimarti Renato ...	2021
The Role of Neurovascular System in Neurodegenerative Diseases.	Molecular neurobiology	Ahmad Ajmal Patel Vanisha Xiao Jianfeng ...	2020

Figure 1: Publication page for the 'Huntington disease' query in our target discovery platform PandaOmics (https://pandaomics.com/).

The use of automatic natural language processing (NLP) methods is imperative for information retrieval (IR) or information extraction (IE) from a large volume of biomedical texts. Several efforts have been made in the past years on entity extraction from scientific publications (Kim et al., 2013; Lee et al., 2016; Allot et al., 2018; Mohan et al., 2018, 2021; Wang and Lo, 2021). For example, Biomedical Entity Search Tool (BEST) uses a dictionary-based indexing strategy to extract ten types of biomedical entities including genes, diseases, drugs, and chemical compounds (Lee et al., 2016), while (Kim et al., 2013; Mohan et al., 2021) adopt machine learning for disease and gene extraction and linking. However, recent works on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) showed that the generalization ability of BERT-based named entity recognition (NER) and entity linking (EL) models is influenced by domain shift or whether the test en-

[1] https://pubmed.ncbi.nlm.nih.gov

Proceedings of EMNLP 2022 Industry Track, pages 606–615
December 9–11, 2020. ©2022 Association for Computational Linguistics

tity/term has been seen in the training set (Miftahutdinov et al., 2020; Tutubalina et al., 2020; Kim and Kang, 2022). Recently, (Soni and Roberts, 2021) compared two commercial search engines with academic prototypes evaluated in the TREC-COVID challenge (Roberts et al., 2020; Voorhees et al., 2021). Their evaluation showed that commercial search engines from Amazon (CORD-19 Search) and Google (COVID-19 Research Explorer) fail to outperform decades-old IR approaches. In particular, the best run (from sabir) was achieved by a SMART system (Buckley, 1985) and used no machine learning or biomedical knowledge. A similar observation has been made for general-domain information retrieval (Thakur et al., 2021), where more efficient approaches e.g. based on dense or sparse embeddings can substantially underperform traditional lexical models like BM25 (Robertson and Zaragoza, 2009).

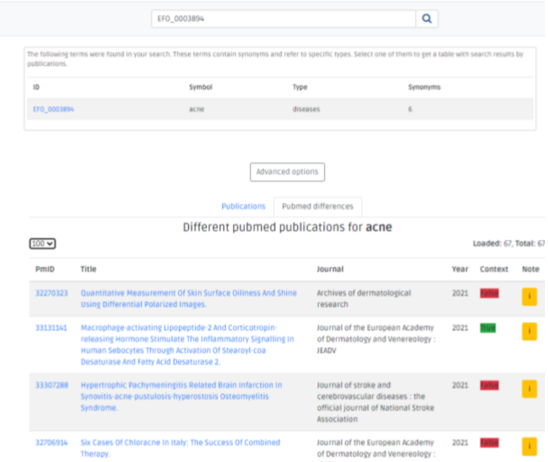
In this paper, we describe the design and evaluation of a BERT-based IE system as an entity-centric search engine for a target discovery platform PandaOmics[2]. In particular, we seek to answer the following research question: considering near excellent performance on NER and EL (Miftahutdinov et al., 2021; Lee et al., 2019), are there models capable of finding relevant publications for disease and gene queries from diverse biomedical subdomains as real-world applications? To help answer this question, we develop a novel search collection of PubMed abstracts for disease and gene queries with corresponding relevance judgments. We evaluate the IE pipeline with two trained BERT-based models for NER and EL and standard document retrieval model BM25 with off-the-shelf Elasticsearch software. We perform error analysis on the models' predictions to shed light on future work directions.

2 Dataset

This section describes our dataset, including queries, and the process of collecting relevance assessments. Table 1 shows statistics of our dataset.

2.1 Queries

In our target discovery platform PandaOmics, a user can enter a gene name or gene symbol like 'PSEN1' (ENSG00000080815) and retrieve all relevant publications and the associated diseases in-



The following terms were found in your search. These terms contain synonyms and refer to specific types. Select one of them to get a table with search results by publications.

ID	Symbol	Type	Synonyms
EFO_0003894	acne	diseases	6

Different pubmed publications for **acne**

Loaded: 67, Total: 67

PmID	Title	Journal	Year	Context	Note
32270323	Quantitative Measurement Of Skin Surface Oiliness And Shine Using Differential Polarized Images.	Archives of dermatological research	2021	False	i
33131141	Macrophage-activating Lipopeptide-2 And Corticotropin-releasing Hormone Stimulate The Inflammatory Signalling In Human Sebocytes Through Activation Of Stearoyl-coa Desaturase And Fatty Acid Desaturase 2.	Journal of the European Academy of Dermatology and Venereology : JEADV	2021	True	i
33307288	Hypertrophic Pachymeningitis Related Brain Infarction In Synovitis-acne-pustulosis-hyperostosis Osteomyelitis Syndrome.	Journal of stroke and cerebrovascular diseases : the official journal of National Stroke Association	2021	False	i
32706914	Six Cases Of Chloracne In Italy: The Success Of Combined Therapy.	Journal of the European Academy of Dermatology and Venereology :	2021	False	i

Figure 2: Task design in our in-house annotation tool with search by disease concept identifier. An annotator selects an abstract and choose one of three labels (relevant/true (green), nonrelevant/false (red), or doubtful (yellow).

cluding Alzheimer's disease (EFO:0000249). An autocomplete feature displays suggestions from disease or gene dictionaries as user search terms. Conversely, the user can enter the disease name 'Alzheimer's disease' to retrieve publications for this concept and the associated targets. These associations are relying on Omics datasets and on a collection of AI-based scores that are based on molecular data and previously published text-based data (see (Ozerov et al., 2016) for more details). As a disease terminology source, we use an internal knowledge base that contains 15,051 concept unique identifiers (CUIs) based on an experimental factor ontology (EFO)[3] (Malone et al., 2010). As a gene terminology source, we use an an internal knowledge base with 28,227 CUIs from Ensembl (Hubbard et al., 2002). We recall that each concept consists of atoms (concept names); all of the atoms within a concept are synonymous (NLM, 2016). As test queries for our dataset, we use the most frequent queries from the platform's logs. These queries are disease CUIs and gene CUIs. In addition, our annotators selected a list of concepts with at least one ambiguous concept name (see Table 2 for examples).

2.2 Relevance Assessments

2.2.1 Pooling

Following standard practice of IR collection building, we employ a *pooling* approach (Lipani et al.,

[2] https://pandaomics.com/

[3] https://www.ebi.ac.uk/efo/

Subset	#queries	avg. number of texts per query		
		relevant label	nonrelevant label	doubtful label
Disease CUI	73	94.86	63.57	9.78
Gene CUI	79	109.39	21.62	5.93
Ambiguous	27	45.94	11.58	0.53
Total	152	102.41	41.76	7.78

Table 1: Summary of statistics of the proposed dataset.

CUI	Ambiguous concept name	Term	Reason	Comment
EFO_0000341	coad	chronic obstructive pulmonary disease	same_synonyms	abbreviation refers to another disease: colon adenocarcinoma (COAD)
EFO_1001998	crps	complex regional pain syndrome	same_synonyms	abbreviation refers to another disease: 'Colorectal polyps'
EFO_1001998	crps	complex regional pain syndrome	same_synonyms	abbreviation refers to another disease: 'chronic regional pain syndrome'
EFO_0000341	dops	chronic obstructive pulmonary disease	refers_to_another	abbreviation refers to another term: direct observation of procedural skills (DOPS)
EFO_0002508	parkinson's disease	parkinson's disease	refers_to_another	author's surname
ENSG00000170345	fos	fos	refers_to_another	refers to fosfomycin

Table 2: A sample of concepts with at least one ambiguous concept name.

2016; Lipani, 2016; Hasibi et al., 2017; Thakur et al., 2021), and combine retrieval results from two main sources:

1. we obtained retrieval results from Elasticsearch; see Sect. 3.2 for the description of this system. Results are pooled from these runs up to depth 100.

2. we obtained retrieval results from PubMed. Results are pooled from these runs up to depth 100, excluding abstracts from the first system.

The final assessment pool contains 23,099 query-abstract pairs (152 abstracts per query on average).

2.2.2 Collecting Relevance Judgments

For each query-abstract pair, we collected the relevance judgments by 2 annotators with biomedical degrees using an in-house annotation tool (Fig. 2). An expert annotator with Ph.D. in biology created a list of queries from logs of our target discovery platform PandaOmics. All annotators are paid biologists in the company. An expert annotator wrote annotation guidelines and educated annotators.

Each annotator selected a disease or gene query from the list of selected identifiers, an abstract with information about the publication year and journal. Abstracts were presented in random order. Annotators were then asked to: (i) judge relevance on a 3-point scale: "relevant", "nonrelevant", or "doubtful", and (ii) categorize the reason for relevance/nonrelevance.

We note that annotators were asked to consider EFO hierarchy during relevance annotation for disease queries. According to the annotation guidelines, only the synonyms belonging to the required level of the hierarchy are relevant. Those terms that are higher in the hierarchy are "wider terms", and those that are lower represent a "narrower case". E.g., while annotating a text for the "prostate adenocarcinoma" query, "prostate cancers" is wider than the term of interest; for the "prostate cancer" query, the "prostate adenocarcinoma" is narrower than the term of interest. Further, we provide a summary of guidelines illustrated with examples.

Relevance The publication relevance to a gene/disease can be determined as true when the gene/disease of interest (its main name or any synonym) is present in the same meaning in an abstract. The term in the abstracts should belong to a disease/gene ontology (and not to any other category,

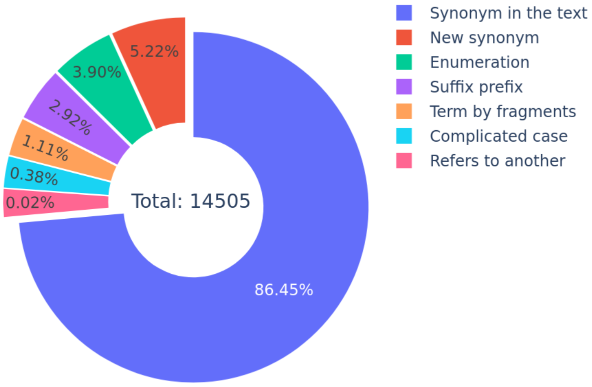


Figure 3: Statistics of Relevance reasons.

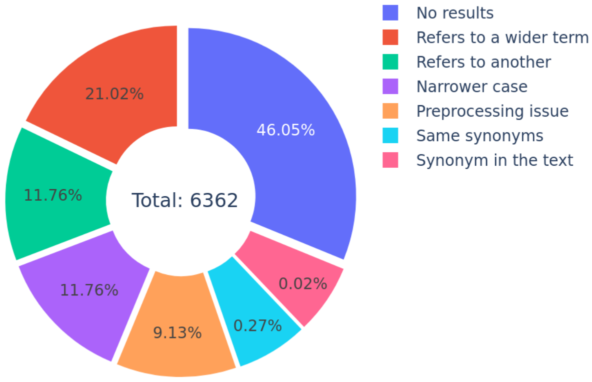


Figure 4: Statistics of Nonrelevance reasons.

e.g. name of a clinical trial, institution, foundation, etc). In particular, there are six reasons for relevance:

1. **synonym in text** – one of the synonyms is precisely present in an abstract;

2. **new synonym** – new synonym for the term of interest, which is absent in our synonyms list, was found;

3. **term by fragments** - an entity is annotated by several fragments of text if: (i) a term is either from the disease of gene ontology; (ii) both fragments are in the same sentence; (iii) the parts of the term are logically connected (according to the author's logic). E.g., the text "...secondary *diabetic* complications, such as *retinopathy*, neuropathy, and nephropathy" (pmid 33109031) should be annotated as TRUE for "diabetic retinopathy";

4. **enumeration** - an entity is annotated by fragments which are separated only with punctuation marks or conjunctions. E.g., the text "asthma-wheezing" (pmid 33276583) should be annotated as a true for both "asthma" and "wheezing", while "AKT1-mTORC1 Axis" (pmid 32404972) should be annotated as TRUE for "AKT1";

5. **suffix/prefix** - an entity was annotated as a part of a word with a suffix/prefix. E.g., we annotate "obesity-induced NAFLD" as a match for "obesity" and add "-induced" as a suffix (we note that there is no "obesity-induced NAFLD" term in the ontology);

6. **complicated case** – a term is encountered in abstract by fragments separated in different sentences, and there is a logical link between them.

Detailed distribution of relevance reasons are given in Fig. 3.

Nonrelevance Nonrelevance of a gene/disease is determined as either no link between the gene/disease and a publication abstract or a wrongly identified relation. The first means the gene/disease is not mentioned in an abstract. The second means that gene/disease is incorrectly linked to an abstract because of one of the following six reasons:

1. **no results** – no results for the term of interest were found in a publication;

2. **refers to another** – gene/disease name (or its abbreviation) is a synonym of some other term, or has some other meanings, which are outside of the ontology (e.g., abbreviation COAD for colon adenocarcinoma refers to another term "anaerobic co-digestion (co-AD)", an abbreviation for Non-alcoholic steatohepatitis refers to another term "Nash equilibria");

3. gene/disease name (or its abbreviation) refers to another term within the ontology (gives *collisions*) because of: (i) **same synonyms** (e.g., abbreviation COAD for Chronic obstructive pulmonary disease refers to another disease "colon adenocarcinoma"); (ii) **refers to a wider term** – publication abstract was found by a wider disease term, which refers not only to a disease of interest, and may give additional non-relevant results (e.g, colon cancer is wider term for colon adenocarcinoma); (iii) **narrower case** – publication abstract was

found by more specific term (e.g., Alzheimer's disease is a narrower case for neurodegenerative disease); (iv) **preprocessing issue** - either ignored punctuation mark ("*background*: *retinopathy*", "*ER-breast cancer*") or is a part of a longer term ("Non-*small cell lung carcinoma*", "Traf2- and *Nck-interacting kinase*").

Detailed distribution of nonrelevance reasons is given in Fig. 4.

We note that our definition of nonrelevance differs from Pubmed search primarily because of the consideration of the concept hierarchy. PubMed search uses Best Match (Fiorini et al., 2018) trained on the user-click information from PubMed search logs. We believe that distinguishing more narrow concepts from broader ones is crucial for target discovery objectives.

Doubtful This category includes publications that mention disease/gene of interest only in keywords/MeSH terms **without an abstract match**. PubMed articles are manually associated with author keywords and MeSH (Medical Subject Headings) (Lipscomb, 2000) as standardized keywords. The reasons for this label are the same as for the relevance label with **synonym in MeSH/keywords** and excluding the "complicated case" category. In 97.65% and 1.6% cases, the annotator associated texts with the synonym in MeSH/keywords and new synonym reasons, respectively.

In 91% and 80% of pairs, two annotators agreed on a relevance label and decision reasons, respectively. When annotators disagreed, the expert annotator was asked to decide whether the relevance labels among with reasons selected by one of the annotators were in fact correct. After this procedure, we obtained the dataset for entity search with 73 disease queries, 79 gene queries, and 23,099 annotated query-entity pairs.

3 Models

The goal of our work is to evaluate retrieval models in a zero-shot setup, with no available training data to train the IR system.

3.1 BERT-based IE pipeline

In our work, we have focused on the extraction of two entity types: disease and gene. Though, we design our IE system with the simplicity of scaling to new entities in mind. The system consists of pipelines, each for a different entity type. The pipelines incorporate two sub-modules: (i) NER sub-module; (ii) EL sub-module. These submodules are applied successively. The first one extracts entities of interest the second one links extracted entities with concepts from given knowledge bases. Taken all together it means that the processing of different types of entities is independent and could be trained and applied separately. As a pretrained transformer model, we use BioBERT base v1.1. (Lee et al., 2019).

Named Entity Recognition In this paper, for reproducibility reasons, we decided to analyze models trained on publicly available academic datasets. Specifically, we train BioBERT on combination of NCBI and CDR Diseases datasets (Doğan et al., 2014; Li et al., 2016) for disease entities and on DrugProt dataset (Miranda et al., 2021) for gene entities. To join the NCBI and CDR Disease datasets, we utilized predefined train/test subsets and combined the datasets within these splits. Thus, the train part of NCBI was combined with the CDR Disease train sets. A similar procedure was carried out to obtain the test part of the combined dataset. We adopted model training hyper-parameters from (Lee et al., 2019). Our model achieves 88.43% and 90.39% of the F-measure on official test sets of disease and gene entities, respectively.

Entity Linking For linking extracted entities to corresponding concepts from dictionaries, we employ state-of-the-art Drug and disease Interpretation Learning with Biomedical Entity Representation Transformer (DILBERT) (Miftahutdinov et al., 2021). This model is based on metric learning and negative sampling, specifically, triplet constraints. Given an entity mention m, a positive concept name c_g and a negative concept name c_n, triplet loss tunes the network such that the distance between m and c_g is smaller than the distance between m and c_n. Details on overall architecture, configuration, hyperparameter search, and evaluation strategies are presented in (Miftahutdinov et al., 2021). The code is publicly available at https://github.com/insilicomedicine/DILBERT. We note that the advantage of DILBERT architecture is the ability to search for the closest concept in a different terminology without retraining the model (cross-terminology use).

Similar to NER, we train models on publicly available academic datasets: CDR Diseases (Li et al., 2016) and BC2GN Genes (Morgan et al., 2008). The models are evaluated on *refined* test sets without entity overlap between train/test sets from (Tutubalina et al., 2020). These sets are publicly available at https://github.com/insilicomedicine/Fair-Evaluation-BERT. Our model achieves 75.8% and 82.4% of accuracy on the refined test sets of diseases and genes, respectively.

Details on models' configurations, speed performance and system deployment are presented in Appendices A and B.

3.2 Elasticsearch BM25

We utilized a popular search engine framework Amazon Elasticsearch/OpenSearch Service[4] that uses OpenSearch v.1.0[5]. OpenSearch is a fork of open source Elasticsearch 7.10[6]. OpenSearch uses BM25 (Robertson and Zaragoza, 2009) to calculate relevance scores. BM25 is a commonly-used bag-of-words retrieval function based on token-matching between two high-dimensional sparse vectors with TF-IDF token weights. We note that (Thakur et al., 2021) recently showed that many approaches with sparse, dense late-interaction architectures outperform BM25 on in-domain evaluation, yet perform poorly on zero-shot setup.

4 Evaluation

For evaluation, we use precision, recall, and F-measure. We calculate the precision as a fraction of relevant documents among all retrieved documents. As well the recall is calculated as a fraction of relevant documents from all possibly relevant documents in the dataset. For experiments, we use query-document pairs with relevant and nonrelevant labels excluding the doubtful category.

Tables 3 and 4 present the performance of the BERT-based pipeline compared to BM25 on the full set of queries and the subset of concept with ambiguous names, respectively. Several observations can be made based on Tables 3 and 4. First, the BERT-based system outperformed BM25 on both sets of the dataset and both types of entities. As expected, the performance difference between the two models is larger on the subset with ambiguous concept names. Third, for the BERT-based pipeline, precision is higher than recall.

Model	**P**	**R**	**F**
Queries with Disease CUIs			
BERT-based	93.97	84.41	88.93
Elasticsearch BM25	82.19	83.33	82.76
Genes			
BERT-based	92.24	85.45	88.71
Elasticsearch BM25	89.92	79.93	84.63
Both			
BERT-based	92.99	84.99	88.81
Elasticsearch BM25	86.23	81.44	83.77

Table 3: IR metrics on the full set of queries.

Model	**P**	**R**	**F**
Queries with Disease CUIs			
BERT-based	97.72	93.81	95.73
Elasticsearch BM25	75.67	96.72	84.91
Genes			
BERT-based	93.02	93.85	93.43
Elasticsearch BM25	79.58	68.88	73.85
Both			
BERT-based	94.9	93.83	94.37
Elasticsearch BM25	77.59	80.39	78.96

Table 4: IR metrics on the subset of queries with ambiguous concepts.

In addition, we investigate search precision further by developing a dataset for out-of-domain abstract detection. Approximately 30,000 records are included in the PubMed journal list. These journals publish papers not only about biological entities, but also on cultural topics, economics and econometrics, artificial intelligence, law, linguistics and language, and so on (out-of-domain categories for us). Our expert annotator manually selected out-of-domain journals on which we expect the IE system to return *zero results*. We randomly select 58,790 abstracts from these journals, where each abstract includes at least one gene of disease concept retrieved by Elasticsearch. In 90% of these abstracts, the BERT-based system did not find any entities.

Error Analysis For error analysis of the BERT-based IE system, we reviewed a sample of 152 false positive (FP) documents and 168 false negative (FN) results. Table 5 provides summary on error categories for FPs. As shown in Table 5, the most frequent category of errors (58%) is related to the ontology hierarchy. Wider cases can also be attributed to a gene when the gene family is mentioned (e.g., Akt (there are Akt1/2/3), ERK

[4] https://aws.amazon.com/opensearch-service/
[5] https://opensearch.org/
[6] https://www.elastic.co/

Reason	N	%
wider term	88	58
refers to another	34	22
synonym in MeSH/keywords	17	11
same synonym	11	7
preprocessing issue	1	1
synonym in the text	1	1

Table 5: Error analysis of IR results on the false positive sample (152 texts).

Reason	Model	N
not found	-	48
largest text span exists	NER	23
not recognized	NER	6
abbreviation	NER	5
wrong recognition	NER	1
wrong mapping	EL	2
largest text span rule/wrong mapping	NER/EL	15

Table 6: Error analysis of NER and EL predictions on the false negative (FN) sample (100 texts).

(there are ERK1/2)). For FNs, 60% of errors (100 abstracts) fell into the synonym in the text category. These documents were additionally analyzed to detect which model (NER or EL) predicted incorrectly (see Table 6). As shown in Table 6, in 23% cases, the NER model predicts a shorter entity which is also known as a boundary problem. E.g., in the text "external validation of the Nonalcoholic [Steatohepatitis]$_{predicted}$ Scoring System in patients" (pmid 33248101) Nonalcoholic Steatohepatitis was mapped to just Steatohepatitis due to NER predictions. Mapping errors are often related to the presence of numbers in gene names or abbreviations. E.g., in a text "orphan nuclear receptor [Nr4a1] mediates perinatal neuroinflammation" (pmid 32606386) entity *Nr4a1* mapped to the Nr4a2 gene instead. For FPs, we additionally analyze 22% of errors (34 abstracts) from the refers to another category. The NER and EL models cause errors in 16 and 11 documents, respectively.

5 Conclusion and Future Work

In this work, we present a comprehensive evaluation of a biomedical entity-centric search engine based on BERT models for disease and gene extraction and linking. This engine is a part of a target discovery platform, where users can return a list of relevant publications given a disease or gene concept query. We evaluate BERT models on two information extraction tasks, entity-centric information retrieval, and out-of-domain abstract detection. Moreover, we present an error analysis for both retrieval and extraction tasks.

This work suggests several interesting directions for future research. We plan to conduct similar studies on other text sources such as full publication texts and patents. Moreover, we plan to expand the list of entity types with pathways and biological processes. To extract explicit associations between drug targets and diseases, we plan to add relation extraction/event detection models and study knowledge graph completion with novel disease-gene edges.

6 Ethics Statement

We outline potential ethical issues with our work below. First, our work focuses on a comprehensive evaluation of the information extraction pipeline for retrieval of relevant scientific texts given queries of disease and gene concepts. Consequently, the developed BERT-based models could reflect many domain-specific biases exhibited by language models. For example, (Sung et al., 2021) showed that predictions on factual triples tend to be highly biased towards a few objects (e.g., "headache", "pain", or "ESR1"). Since pretrained language models are used for initialization, it is possible to reflect biased patterns in open-world applications. Second, our NLP engine is a part of the target discovery platform `PandaOmics` which intend to identify targets (genes/proteins) through deep feature selection, causality inference, and de novo pathway reconstruction (Ozerov et al., 2016). We use the NLP engine to assess the targets' novelty and disease association via the analysis of research publications. The imperfect completeness of the extracted information can be especially reflected in the small number of publications in the search results about rare diseases, making it difficult for subsequent analysis. Third, we use EFO and Ensembl as primary resources with disease hierarchy and concepts' synonyms. For example, (Miftahutdinov et al., 2021) demonstrated that degradation in the accuracy from the full disease dictionary to a 30% of the dictionary is significant for disease linking in clinical trials. Moreover, consistent description of these entities has numerous differing standards and opportune incorporation of new human disease terms and targets is still necessary.

References

Alexis Allot, Yifan Peng, Chih-Hsuan Wei, Kyubum Lee, Lon Phan, and Zhiyong Lu. 2018. Litvar: a semantic search engine for linking genomic variant data in pubmed and pmc. *Nucleic acids research*, 46(W1):W530–W536.

Chris Buckley. 1985. Implementation of the smart information retrieval system. Technical report, Cornell University.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers), Minneapolis, USA*, pages 4171–4186.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Nicolas Fiorini, Kathi Canese, Grisha Starchenko, Evgeny Kireev, Won Kim, Vadim Miller, Maxim Osipov, Michael Kholodov, Rafis Ismagilov, Sunil Mohan, et al. 2018. Best match: new relevance search for pubmed. *PLoS biology*, 16(8):e2005343.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 267–274.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: a test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1265–1268.

Tim Hubbard, Daniel Barker, Ewan Birney, Graham Cameron, Yuan Chen, L Clark, Tony Cox, J Cuff, Val Curwen, Thomas Down, et al. 2002. The ensembl genome database project. *Nucleic acids research*, 30(1):38–41.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

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

Jeongkyun Kim, Seongeun So, Hee-Jin Lee, Jong C. Park, Jung-jae Kim, and Hyunju Lee. 2013. DigSee: disease gene search engine with evidence sentences (version cancer). *Nucleic Acids Research*, 41(W1):W510–W517.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Sunwon Lee, Donghyeon Kim, Kyubum Lee, Jaehoon Choi, Seongsoon Kim, Minji Jeon, Sangrak Lim, Donghee Choi, Sunkyu Kim, Aik-Choon Tan, et al. 2016. Best: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PloS one*, 11(10):e0164680.

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. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Aldo Lipani. 2016. Fairness in information retrieval. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 1171, New York, NY, USA. Association for Computing Machinery.

Aldo Lipani, Mihai Lupu, and Allan Hanbury. 2016. The curious incidence of bias corrections in the pool. In *European Conference on Information Retrieval*, pages 267–279. Springer.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. 2010. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118.

Zulfat Miftahutdinov, Ilseyar Alimova, and Elena Tutubalina. 2020. On biomedical named entity recognition: experiments in interlingual transfer for clinical and social media texts. In *European Conference on Information Retrieval*, pages 281–288. Springer.

Zulfat Miftahutdinov, Artur Kadurin, Roman Kudrin, and Elena Tutubalina. 2021. Medical concept normalization in clinical trials with drug and disease representation learning. *Bioinformatics*, 37(21):3856–3864.

Antonio Miranda, Farrokh Mehryary, Jouni Luoma, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. 2021. Overview of drugprot biocreative vii track: quality evaluation and large scale text mining of drug-gene/protein relations. In *Proceedings of the seventh BioCreative challenge evaluation workshop*.

Sunil Mohan, Rico Angell, Nicholas Monath, and Andrew McCallum. 2021. Low resource recognition and linking of biomedical concepts from a large ontology. In *Proceedings of the 12th ACM conference on bioinformatics, computational biology, and health informatics*, pages 1–10.

Sunil Mohan, Nicolas Fiorini, Sun Kim, and Zhiyong Lu. 2018. A fast deep learning model for textual relevance in biomedical information retrieval. In *Proceedings of the 2018 World Wide Web Conference*, pages 77–86.

Alexander A Morgan, Zhiyong Lu, Xinglong Wang, Aaron M Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jörg Hakenberg, et al. 2008. Overview of biocreative ii gene normalization. *Genome biology*, 9(S2):S3.

NLM. 2016. Umls glossary.

Ivan V Ozerov, Ksenia V Lezhnina, Evgeny Izumchenko, Artem V Artemov, Sergey Medintsev, Quentin Vanhaelen, Alexander Aliper, Jan Vijg, Andreyan N Osipov, Ivan Labat, et al. 2016. In silico pathway activation network decomposition analysis (ipanda) as a method for biomarker development. *Nature communications*, 7(1):1–11.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2020. Trec-covid: rationale and structure of an information retrieval shared task for covid-19. *Journal of the American Medical Informatics Association*, 27(9):1431–1436.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Sarvesh Soni and Kirk Roberts. 2021. An evaluation of two commercial deep learning-based information retrieval systems for covid-19 literature. *Journal of the American Medical Informatics Association*, 28(1):132–137.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Elena Tutubalina, Artur Kadurin, and Zulfat Miftahutdinov. 2020. Fair evaluation in concept normalization: a large-scale comparative analysis for bert-based models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6710–6716.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

Lucy Lu Wang and Kyle Lo. 2021. Text mining approaches for dealing with the rapidly expanding literature on covid-19. *Briefings in Bioinformatics*, 22(2):781–799.

Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th international conference on world wide web*, pages 1271–1279.

A System deployment

Our system is packaged in a docker container, which is run by schedule. Since the container is self-contained any scheduler could be used. The pipeline of documents processing in service is as follows: (i) load to the local storage previously unlabeled documents from a database (ii) extract and link entities from documents using the BERT-based pipeline (iii) upload the labeled documents to the database. We store our documents in MongoDB (`https://www.mongodb.com`). The service is implemented substantially on python, with entrypoints written in shell. To load/upload documents from/toMongoDB we use PyMongo library (`https://pymongo.readthedocs.io`). After the labeled documents are loaded to the MongoDB we utilize Elasticsearch as a search index. Customers of the drug discovery platform send concept CUI as a query, afterward, the backend retrieve all documents containing specified CUI and transfer them to the frontend.

B Configuration details and speed performance

For NER and EL encoders, we apply the fine-tuned on downstream task BioBERT v1.1 with 12 heads, 12 layers, 768 hidden units per layer, and a total of 110M parameters. We train our NER model using AdamW (Loshchilov and Hutter, 2018) optimizer for 20 epochs with a batch size equal to 48 and learning rate equal to 5e-5. The EL model is trained with the same optimizer and learning rate for 5 epochs and batch size equal to 32. At the inference and training time, we restrict the length of the sequence up to 128 sub-tokens for entity recognition and up to 28 sub-tokens for linking.

For NER sub-module we use Huggingface python library (https://huggingface.co), for EL we apply sentence-transformers library (https://www.sbert.net). At the inference time, the EL model uses the FAISS library (Johnson et al., 2019) with GPU support for a fast nearest neighbor search by comparing vectors with Euclidean distance. Embeddings of all terminologies' concepts are indexed.

We note that deployed models are trained on in-house datasets with similar parameters and evaluation metrics that are not publicly available due to company policy.

We profiled retrieval speed on a server with Intel Xeon CPU E5-2660 2.00GHz and 256GB memory. First, we precomputed all embeddings for all concepts (500 thousand). On a single Nvidia TITAN X GPU, it takes about 7 minutes to compute all embeddings. Given that all embeddings are indexed on Nvidia TITAN X GPU using IndexFlatL2 index type 5 thousand documents processing takes 390 seconds, which is 0.08 seconds per document. Most of this time, specifically 359 seconds, is taken by the NER sub-module.