Learning from Negative Samples in Biomedical Generative Entity Linking

Chanhwi Kim[1*], Hyunjae Kim[1*], Sihyeon Park[1], Jiwoo Lee[1], Mujeen Sung[2†], Jaewoo Kang[1,3†]

[1]Korea University, [2]Kyung Hee University, [3]AIGEN Sciences
{chanhwi_kim, hyunjae-kim, sh10, hijiwoo7}@korea.ac.kr
mujeensung@khu.ac.kr, kangj@korea.ac.kr

Abstract

Generative models have become widely used in biomedical entity linking (BioEL) due to their excellent performance and efficient memory usage. However, these models are usually trained only with positive samples, i.e., entities that match the input mention's identifier, and do not explicitly learn from hard negative samples, which are entities that look similar but have different meanings. To address this limitation, we introduce ANGEL (Learning from Negative Samples in Biomedical Generative Entity Linking), the first framework that trains generative BioEL models using negative samples. Specifically, a generative model is initially trained to generate positive entity names from the knowledge base for given input entities. Subsequently, both correct and incorrect outputs are gathered from the model's top-k predictions. Finally, the model is updated to prioritize the correct predictions through preference optimization. Our models outperform the previous best baseline models by up to an average top-1 accuracy of 1.4% on five benchmarks. When incorporating our framework into pre-training, the performance improvement increases further to 1.7%, demonstrating its effectiveness in both the pre-training and fine-tuning stages. The code and model weights are available at https://github.com/dmis-lab/ANGEL.

1 Introduction

Biomedical entity linking (BioEL) involves aligning entity mentions in text with standardized concepts from biomedical knowledge bases (KB) such as UMLS (Bodenreider, 2004) or MeSH (Lipscomb, 2000).[1] BioEL encounters significant challenges due to the diverse and ambiguous nature of biomedical terminology, including synonyms, abbreviations, and terms that look similar but have different meanings. For instance, 'ADHD' (CUI:*C1263846*, where CUI stands for Concept Unique ID) has synonyms such as hyperkinetic disorder and attention deficit hyperactivity disorder. Additionally, 'ADA' can be mapped to either adenosine deaminase (CUI:*C1412179*) or American Diabetes Association (CUI:*C1705019*) depending on the context in which the entity appears.

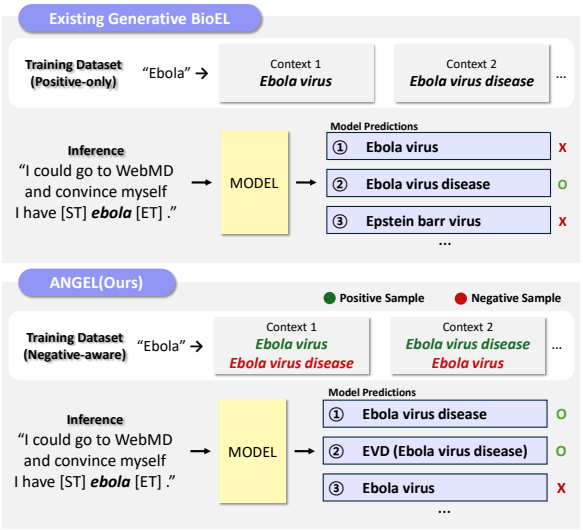

Figure 1: Comparison of training approaches between existing generative BioEL models and our ANGEL method. The main limitation of current generative BioEL methods is that they are trained only on positive samples. This restricts their ability to distinguish between entity names that are similar in surface form but different in meaning depending on the context. Our ANGEL framework addresses this issue by training the model to prefer positive samples over negative ones.

Recent studies have focused on addressing these challenges, broadly categorized into two approaches: similarity-based and generative BioEL. Similarity-based models (Sung et al., 2020; Liu et al., 2021; Lai et al., 2021; Bhowmik et al., 2021) encode input mentions and entities from KBs into the same vector space using embedding models.

[*] Co-first authors; [†] Co-corresponding authors

[1]UMLS and MeSH are short for the Unified Medical Language System and Medical Subject Headings, respectively.

Findings of the Association for Computational Linguistics: ACL 2025, pages 10714–10730

They then calculate similarity scores to identify the most similar entities for each input entity. Although these approaches have achieved remarkable improvements, they require significant space to index and load embedding vectors for all candidate entities (De Cao et al., 2020). Furthermore, representing both the input and candidate entities as single vectors using a bi-encoder can limit the quality of their representations, making it difficult to handle challenging cases.

On the other hand, generative models (De Cao et al., 2020; Yuan et al., 2022a,b), built upon an encoder-decoder structure (Lewis et al., 2020; Raffel et al., 2020), directly generate the most likely entity name from the KB for the input entity. The output space is dynamically controlled through a constrained decoding strategy, ensuring that only entities from the target KB are generated. Generative models offer several advantages over similarity-based models, including greater memory efficiency and higher performance. They eliminate the need to index large external embedding vectors, and their auto-regressive formulation effectively cross-encodes the input document and candidate entities.

However, existing generative models are trained solely on positive samples and do not explicitly learn from negative samples. Despite their high performance, they encounter limitations when distinguishing between biomedical entities with similar surface forms but different meanings. Although similarity-based models address this issue by incorporating negative samples through synonym marginalization (Sung et al., 2020) or contrastive learning (Liu et al., 2021), applying these approaches to generative models is not straightforward. Consequently, generative models may overfit to surface-level features, reducing the models' ability to generalize effectively across varied contexts, as illustrated in Figure 1.

To harness the benefits of generative approaches while overcoming their limitation of not using negative samples, we introduce a novel training framework, ANGEL. Our framework operates in two stages: positive-only training and negative-aware training (see Figure 2). In the first stage, a generative model is trained to generate biomedical terms from the KB that share the same identifier as the given input entity. In the second stage, we gather both correct and incorrect outputs from the model's top-k predictions. The model is then updated to prioritize the correct predictions using a preference optimization algorithm (Burges, 2010; Rafailov et al., 2024). Models trained on our ANGEL framework significantly outperform the previous best similarity-based and generative BioEL models, achieving an average accuracy improvement of 1.7% across five datasets. Our contributions are as follows:

- We introduce ANGEL, the first-of-its-kind training framework that utilizes negative samples in generative entity linking. ANGEL overcomes the limitations of existing generative approaches by effectively employing negative samples during training.
- ANGEL is a versatile framework, demonstrating its applicability in both the pre-training and fine-tuning phases, leading to performance improvements at each stage. Additionally, our method is model-agnostic, consistently improving results across various backbone language models, with gains ranging from 0.9% to 1.7%.
- Our best model, pre-trained and fine-tuned with our framework, outperforms the previous best baseline model by 1.7% across five benchmark datasets.

2 Related Work

2.1 Biomedical Entity Linking

Biomedical entity linking (BioEL), also known as biomedical entity normalization, is a crucial task because of its application in several downstream tasks in the biomedical domain, such as literature search (Lee et al., 2016), knowledge extraction (Li et al., 2016a; Xiang et al., 2021; Zhang et al., 2023), knowledge graph alignment (Cohen and Hersh, 2005; Lin et al., 2022), and automatic diagnosis (Shi et al., 2021; Yuan and Yu, 2024). Typically, it is assumed that the target mention is already provided, and the task is solely to link this mention to the appropriate entity name from the KB. End-to-end BioEL (Zhou et al., 2021; Ujiie et al., 2021), which also involves identifying mentions within a sentence, is being actively researched, but this is not our focus and will not be discussed in detail.

Traditional classification-based approaches (Limsopatham and Collier, 2016a; Miftahutdinov et al., 2019) employed a softmax layer for classification, treating concepts as categorical variables and thereby losing the detailed information of concept names. Similarity-based (Sung et al.,

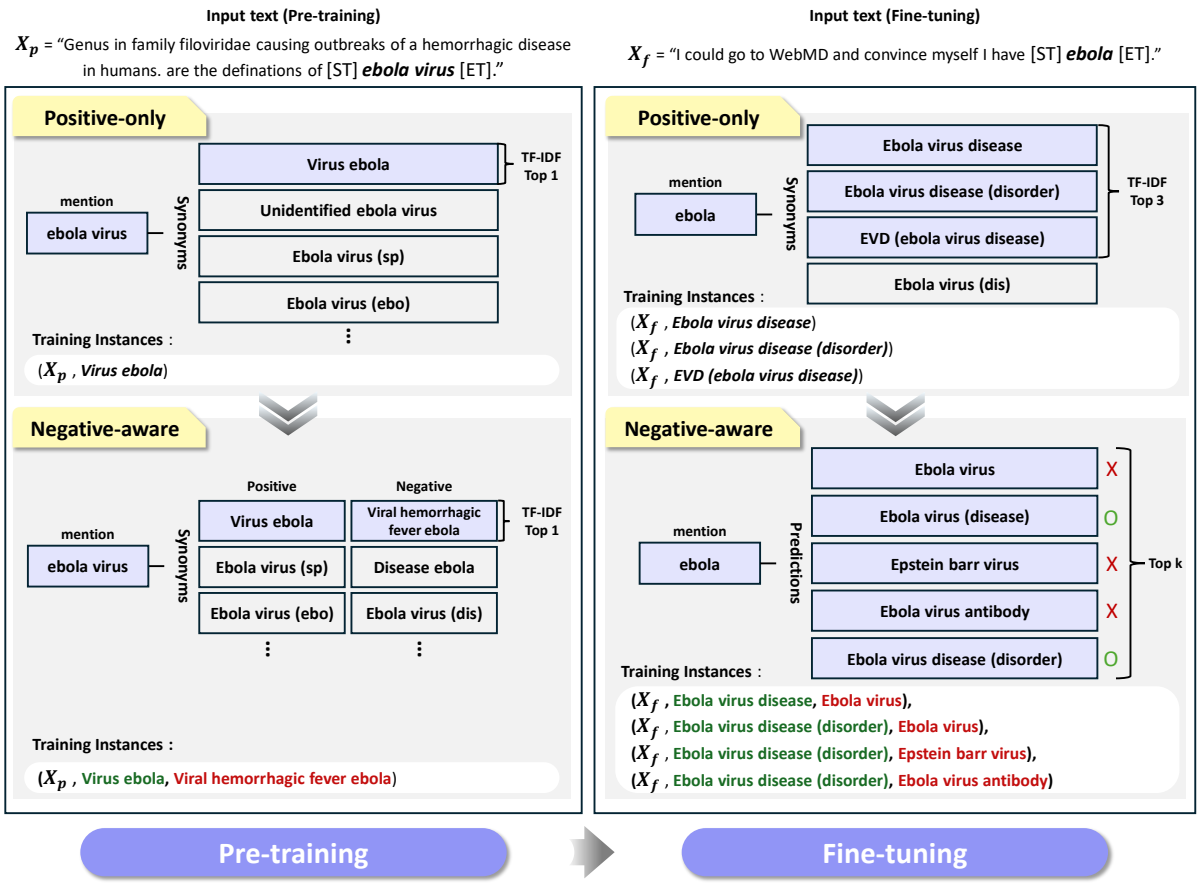

Figure 2: Overview of our method ANGEL. The core idea is to enhance both pre-training and fine-tuning by incorporating negative samples, which are obtained either through TF-IDF similarity or the model's top-k predictions. This approach helps the model distinguish subtle differences between correct and incorrect entities.

2020; Liu et al., 2021; Lai et al., 2021; Zhang et al., 2022) models have significantly improved BioEL performance, which encodes mentions and candidate entity names in the same vector space. They are characterized by high memory consumption due to the need to encode entities into pre-computed embeddings, posing scalability challenges with large datasets (De Cao et al., 2020). Several studies have integrated the concept of clustering into BioEL (Angell et al., 2021; Agarwal et al., 2022).

2.2 Generative Entity Linking

Generative models have become a powerful method for entity linking by overcoming the limitations of similarity-based models. The GENRE framework (De Cao et al., 2020) was the first to demonstrate this approach. To enhance precision and reduce memory usage, GENRE introduced a constrained decoding method (Hokamp and Liu, 2017) using a prefix tree (trie), which restricts the output space to valid entity names. This technique also facilitates easy updates to the set of entities, making the system highly adaptable to changes in the KB. In the biomedical field, notable examples of generative models include GenBioEL (Yuan et al., 2022b) and BioBART (Yuan et al., 2022a). GenBioEL, in particular, is the first model to apply a generative model BART (Lewis et al., 2020) to BioEL, after pre-training it using UMLS. Additionally, several hybrid approaches, known as retrieve-and-generate methods, have been proposed (Xu et al., 2023; Lin et al., 2024). In these methods, a similarity-based model first retrieves the top-k candidates, which are then reranked using a generative model. Although generative approaches have shown high performance, their training has typically been limited to positive samples, as discussed in the introduction section. This absence of explicit negative sample learning often leads to confusion when entities share very similar surface forms but represent different concepts. In this study, we introduce the use of negative samples during training and demonstrate that this approach can significantly enhance

the performance of generative models.

3 Method

3.1 Task Formulation

Let $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ be a human-labeled dataset, where $\mathbf{x}_n$ represents an input text and y_n is the gold identifier defined in a KB denoted by $\mathcal{E}$. Each $\mathbf{x}_n = (\mathbf{c}_n^-, \mathbf{m}_n, \mathbf{c}_n^+)$ contains a target entity mention $\mathbf{m}_n$ along with its surrounding contextual information $\mathbf{c}_n^-$ and $\mathbf{c}_n^+$, which represents the tokens before and after the entity mention $\mathbf{m}_n$, respectively. For simplicity, we will omit the subscript n. Our goal is to map each mention $\mathbf{m}$ to its corresponding gold identifier y from the set of entity names $\mathcal{E}$. To achieve this, we define the model's prediction y^* as follows:

$$y^* = \mathcal{F}(\text{argmax}_{\mathbf{e}\in\mathcal{E}}\ p_\theta(\mathbf{e}|\mathbf{m})), \quad (1)$$

where $\mathbf{e}$ is an entity name defined in the KB, $\mathcal{F}$ is a function that aligns entities to their identifiers, and θ represents the model parameters.

A single gold identifier may have multiple associated entity names that refer to the same concept; we refer to these as synonyms. Previous generative BioEL approaches train the model to generate a textual synonym $\mathbf{s} \in \mathcal{S}_y$, where $\mathcal{S}_y \subset \mathcal{E}$ denotes the set of entity names associated with the identifier y, in an autoregressive manner as follows:

$$p_\theta(\mathbf{s} \mid \mathbf{x}, \mathbf{v}) = \prod_{t=1}^{T} p_\theta(s_t \mid s_{<t}, \mathbf{x}, \mathbf{v}), \quad (2)$$

where T is the number of tokens of the synonym $\mathbf{s}$, s_t indicates the t-th token of the synonym, and $\mathbf{v}$ is the prompt. In an encoder-decoder model structure (Lewis et al., 2020), the input to the encoder is formatted as follows:

$$\texttt{[BOS]}\ \mathbf{c}^-\ \texttt{[ST]}\ \mathbf{m}\ \texttt{[ET]}\ \mathbf{c}^+\ \texttt{[EOS]},$$

where the special tokens [ST] and [ET] surround the target mention, and the special tokens [BOS] and [EOS] represent the 'Begin Of Sentence' and 'End Of Sentence,' respectively. The prefix prompt $\mathbf{v}$ to the decoder, represented as '$\mathbf{m}$ is', is concatenated with [BOS] and input to the decoder. The prompt is designed to make the decoder's output resemble a natural language sentence, which helps to minimize discrepancies between language modeling and fine-tuning on the BioEL task.

As shown in Equation 2, existing models are trained solely to predict synonyms for the input mention (i.e., positive samples), without leveraging negative samples. In contrast, we propose a novel approach using negative samples, which we will describe in detail in the following sections.

3.2 ANGEL Framework

Our framework comprises two main stages: positive-only training, which warms up the model using positive samples to learn morphological similarities among synonyms, and negative-aware training, which progressively refines the model by incorporating negative samples (see Figure 2).

Positive-only training We initialize the model to generate synonyms, similar to previous methods (see Equation 2). For the input mention, we select the most similar synonyms based on their vector similarity, which is calculated as follows:

$$\hat{\mathcal{S}}_y = \text{argsort}_{\mathbf{s}\in\mathcal{S}_y}(\text{TFIDF}(\mathbf{m}, \mathbf{s})), \quad (3)$$

where $\text{TFIDF}(\cdot)$ returns the TF-IDF similarity between tri-grams of the mention and its synonyms. We use the top-k subset $\hat{\mathcal{S}}_y[:\text{k}] = \{\hat{\mathbf{s}}^{(1)}, \ldots, \hat{\mathbf{s}}^{(\text{k})}\}$ as training instances for each mention.

Negative-aware training After obtaining the top-k predictions from the model for each mention in the training set, we construct a dataset of triplets $(\mathbf{x}, \mathbf{e}_w, \mathbf{e}_l)$, where $\mathbf{x}$ denotes the mention along with its context (if available), $\mathbf{e}_w$ is the correct (preferred) entity, and $\mathbf{e}_l$ is an incorrect (dispreferred) entity. From all possible $(\mathbf{e}_w, \mathbf{e}_l)$ pairs, we retain only those for which the model ranks the incorrect entity $\mathbf{e}_l$ above the correct one $\mathbf{e}_w$, thereby reflecting an incorrect model preference. If the top-ranked prediction is already correct, we pair this correct entity $\mathbf{e}_w$ with the highest-ranked incorrect entity $\mathbf{e}_l$ to preserve the model's original preference structure.

We denote the resulting training set by $\mathcal{D}'$ and fine-tune the model using a pairwise preference loss, formulated as follows:

$$\mathcal{L}(\theta) = -\,\mathbb{E}_{(\mathbf{x},\mathbf{e}_w,\mathbf{e}_l)\sim\mathcal{D}'}[\log\sigma(\beta(r_\theta(\mathbf{e}_w \mid \mathbf{x}) - r_\theta(\mathbf{e}_l \mid \mathbf{x})))], \quad (4)$$

where $r_\theta(\mathbf{e} \mid \mathbf{x})$ is a differentiable scoring function (e.g., $\log p_\theta(\mathbf{e} \mid \mathbf{x})$), σ is the sigmoid function, and β is a temperature (scaling) hyperparameter. A recent instantiation of this general preference learning framework is Direct Preference Optimization (DPO) (Rafailov et al., 2024), where the scoring

function is defined as a log-ratio with respect to a reference model p_{ref} as follows:

$$r_\theta(\mathbf{e} \mid \mathbf{x}) = \log \frac{p_\theta(\mathbf{e} \mid \mathbf{x})}{p_{\text{ref}}(\mathbf{e} \mid \mathbf{x})}, \quad (5)$$

where p_θ is the generative model being trained, p_{ref} is a reference generative model trained in a prior stage using positive-only data. We adopt DPO as a practical instantiation within our framework because it offers a principled and empirically effective way to incorporate prior model behavior through a reference distribution.

Applying ANGEL in pre-training Our framework supports not only fine-tuning with labeled datasets but also pre-training with the KB. Specifically, we automatically generate surrounding contextual information for each entity in the KB, using clause templates or definitions, as outlined in GenBioEL (Yuan et al., 2022b). When the entity definition $\mathbf{d}_y$ corresponding to the identifier y is available, a synonym and its definition are integrated into a pre-defined clause template as follows:[1]

`[BOS] [ST]` $\mathbf{s}$ `[ET] is defined as` $\mathbf{d}_y$ `[EOS].`

When no definitions are available in the KB, we replace $\mathbf{d}_y$ with alternative synonyms as follows:

`[BOS] [ST]` $\mathbf{s}_1$ `[ET] has synonyms such as` $\mathbf{s}_2$ `[EOS],`

where $\mathbf{s}_1$ and $\mathbf{s}_2$ are different synonyms. The input for the decoder is "`[BOS]` $\mathbf{s}$ (or $\mathbf{s}_1$) `is`" and the expected output is another synonym (e.g., $\mathbf{s}_2$) selected from the remaining synonyms.

The pre-training process, like fine-tuning, is divided into two stages: positive-only training and negative-aware training. However, due to the typically large scale of the KB, efficiency considerations are particularly important. In positive-only training, rather than utilizing all possible synonym combinations within the KB, we identify, for each entity, the most similar synonym based on TF-IDF similarity and designate it as the target synonym. For negative-aware training, instead of selecting negatives from the model's predictions, negative samples are selected from entities exhibiting the highest TF-IDF similarity to the input mentions but possessing distinct identifiers.

[1]Refer to Yuan et al. (2022b) for the full set of templates.

4 Experiments

4.1 Datasets

We utilized five popular BioEL benchmark datasets: NCBI-disease (Doğan et al., 2014), BC5CDR (Li et al., 2016b), COMETA (Basaldella et al., 2020), AskAPatient (Limsopatham and Collier, 2016b), and Medmentions (Mohan and Li, 2019), with the ST21pv subset used for Medmentions. Due to the lack of a test set in the AskAPatient dataset, we adhered to the 10-fold evaluation protocol outlined by Limsopatham and Collier (2016b). Also, AskAPatient dataset does not include context for the mentions. In the following tables, NCBI-disease, AskAPatient, and Medmentions are denoted as NCBI, AAP, and MM-ST21pv, respectively. Refer to Appendix A for detailed descriptions and statistics.

4.2 Baseline Models

We used top-performing similarity-based models (Sung et al., 2020; Liu et al., 2021; Lai et al., 2021; Zhang et al., 2022) as our baselines. Notably, Prompt-BioEL (Xu et al., 2023) employs a re-ranking-based approach. In the first stage, a similarity-based model, such as SapBERT, retrieves the top-k candidate entities from the knowledge base. In the second stage, these candidates are reranked using a cross-encoder. Although Prompt-BioEL may not be directly comparable, as it incorporates additional modules on top of existing models, we report its performance alongside for reference. Additionally, we include the previously best-performing generative models for comparison. (1) BART-large (Lewis et al., 2020) is an encoder-decoder language model pre-trained on a general-domain corpus. (2) BioBART-large (Yuan et al., 2022a) is the BART-large model continuously pre-trained on a biomedical-domain corpus. (3) GenBioEL (Yuan et al., 2022b) is initialized with the weights of the BART-large model and then pre-trained specifically for BioEL using UMLS. We excluded several models (Agarwal et al., 2022; Lin et al., 2024) due to the lack of publicly available code or the difficulty in reproducing their reported performance.

4.3 Implementation Details

Our framework was applied to each of these models during fine-tuning, referred to as ANGEL_{FT}, and during both pre-training and fine-tuning, referred to as $\text{ANGEL}_{\text{PT+FT}}$. For pre-training, we uti-

Model	NCBI	BC5CDR	COMETA	AAP	MM-ST21pv	Average
Similarity-based BioEL & Re-ranking						
BioSYN (Sung et al., 2020)	91.1	93.3†	71.3	86.5†	OOM	-
SapBERT (Liu et al., 2021)	92.3	88.6†	75.1	89.0	50.3†	79.1
ResCNN (Lai et al., 2021)	92.4	94.0†	80.1	77.4†	55.0	79.3
KRISSBERT (Zhang et al., 2022)	91.3	72.0†	80.1†	83.1†	72.2	79.7
Prompt-BioEL (Xu et al., 2023)	91.9†	94.3†	82.7†	89.7†	72.6†	86.2
Generative BioEL (reported)						
BART (Lewis et al., 2020)	90.2	92.5	80.7	88.8	71.5	84.7
BioBART (Yuan et al., 2022a)	89.9	93.3	81.8	89.4	71.8	85.2
GenBioEL (Yuan et al., 2022b)	91.9	93.3	81.4	89.3	-	-
Generative BioEL (reproduced)						
BART† (Lewis et al., 2020)	90.3	93.0	80.4	88.7	70.1	84.5
+ ANGEL$_{\text{FT}}$ (**Ours**)	91.4 (+1.1)	93.6 (+0.6)	81.3 (+0.9)	89.5 (+0.8)	71.2 (+1.1)	85.4 (+0.9)
BioBART† (Yuan et al., 2022a)	89.4	93.5	81.3	89.3	71.3	85.0
+ ANGEL$_{\text{FT}}$ (**Ours**)	91.9 (+2.5)	**94.7** (+1.2)	82.2 (+0.9)	89.9 (+0.6)	**73.4** (+2.1)	86.4 (+1.4)
GenBioEL† (Yuan et al., 2022b)	91.0	93.1	80.9	89.3	70.7	85.0
+ ANGEL$_{\text{FT}}$ (**Ours**)	92.5 (+1.5)	94.4 (+1.3)	82.4 (+1.5)	89.9 (+0.6)	71.9 (+1.2)	86.2 (+1.2)
+ ANGEL$_{\text{PT+FT}}$ (**Ours**)	**92.8** (+1.8)	94.5 (+1.4)	**82.8** (+1.9)	**90.2** (+0.9)	73.3 (+2.6)	**86.7** (+1.7)

Table 1: The top-1 accuracy of the models across the five BioEL datasets. Our ANGEL framework is applied to generative BioEL models during fine-tuning (ANGEL$_{\text{FT}}$) and both pre-training and fine-tuning (ANGEL$_{\text{PT+FT}}$). '†': the results have been reproduced. 'OOM': an out-of-memory error occured when using a single 80G A00 GPU.

lized the 2020AA version of the UMLS database,[2] which comprises 3.09M entities, of which 199K concepts contain definitions. During pre-training, we saved checkpoints every 500 steps over 5 epochs and selected the best one based on the validation sets. We used top-3 synonyms as positive samples in positive-only training. The other hyperparameter configurations are detailed in Appendix B. In pre-processing, following Yuan et al. (2022b), we expanded abbreviations using AB3P (Sohn et al., 2008), lowercase texts, mark mention boundaries with special tokens `[ST]` and `[ET]`, and discard mentions that overlap or are missing from the target KB. During pre-training, our models were trained using eight 80G A100 GPUs for 12 hours. During fine-tuning, a single A100 GPU was used.

4.4 Results

Consistent with previous studies (Sung et al., 2020; Liu et al., 2021), we used accuracy at top-1 (Acc@1) as our evaluation metric, which quantifies the percentage of mentions where the model correctly ranks the gold standard identifier as the top choice. To assess statistical significance, we employed bootstrapping with the same sample size as the original datasets, repeating the process 100 times, followed by a paired t-test. Table 1 shows that our framework consistently outperformed the performance of generative models ($p \leq 8.2e^{-22}$ for all comparisons). Specifically, our fine-tuning method (i.e., ANGEL$_{\text{FT}}$) improved the Acc@1 scores of BART, BioBART, and GenBioEL by 0.9%, 1.4%, and 1.2%, respectively. When pre-training is also applied (i.e., ANGEL$_{\text{PT+FT}}$) to GenBioEL, the improvement increases to 1.7%, further highlighting the effectiveness of both pre-training and fine-tuning in ANGEL.

Similarity-based models often exhibit limited robustness, with performance varying significantly across datasets. In contrast, generative models tend to deliver more consistent results, highlighting a key strength. Among the baseline methods, the re-ranking-based model Prompt-BioEL achieves strong performance, substantially outperforming its underlying retriever, SapBERT, though at the cost of increased inference time. Notably, our ANGEL$_{\text{PT+FT}}$ model surpasses Prompt-BioEL across all datasets without relying on any re-ranking component, achieving an average improvement of 0.5%. Given this strong baseline, incorporating a re-ranking component into our model in future work may further enhance performance.

[2] https://www.nlm.nih.gov/research/umls/licensedcontent/umlsarchives04.html

Model	NCBI	BC5CDR	COMETA	AAP	MM-ST21pv	Average
*Models **with** Negative-aware Training*						
ANGEL (**Ours**)	92.8	**94.5**	**82.8**	**90.2**	**73.3**	**86.7**
Prediction-based $\mathbf{e}_l$ ⇒ TF-IDF-based $\mathbf{e}_l$	91.8	94.4	81.6	90.0	71.5	85.9
$p_\theta(\mathbf{e}_l) > p_\theta(\mathbf{e}_w)$ Pairs ⇒ All Possible Pairs	**92.9**	94.0	81.9	90.0	72.0	86.2
$\mathbf{e}_l$ within Top-5 ⇒ Top-10 Predictions	92.5	94.0	82.1	89.6	72.6	86.2
*Models **without** Negative-aware Training*						
GenBioEL (Yuan et al., 2022b)	91.0	93.1	80.9	89.3	70.7	85.0

Table 2: The ablation study on positive ($\mathbf{e}_w$) and negative ($\mathbf{e}_l$) pair selection during negative-aware fine-tuning. '⇒' indicates a modification in our method.

Model	NCBI	BC5CDR
TF-IDF (trigram-based)	**91.0**	**92.6**
BioBERT-NLI	90.1	71.9
SapBERT	90.2	84.1

Table 3: Comparison of similarity models for retrieving positive and negative samples from KBs.

Model	FT	BC5CDR	AAP
BART	✗	0.8	15.6
GenBioEL	✗	33.1	50.6
+ ANGEL (**Ours**)	✗	**49.7**	**61.5**
BART	✓	93.0	88.7
GenBioEL	✓	93.1	89.3
+ ANGEL (**Ours**)	✓	**94.5**	**90.2**

Table 4: The top-1 accuracy of models with different pre-training strategies, along with the fine-tuned scores. 'FT' denotes fine-tuning, with ✗ representing pre-trained models without fine-tuning, and ✓ indicating models fine-tuned on human-annotated training sets.

5 Analysis

5.1 Ablation Study

We conducted in-depth analyses on the selection of positive and negative pairs, the effect of similarity models on synonym retrieval, and the effect of pre-training. Additional analyses and results—including the number of synonyms used in positive-only training and the effect of optimization functions—can be found in Appendix C.

Selection of positive and negative pairs Analyzing the impact of how positive-negative pairs are constructed during negative-aware training is crucial for determining the optimal strategy for selecting hard negatives and the appropriate number of pairs. We investigated the effects of three factors: (1) negative sampling techniques (i.e., whether to use the model's incorrect predictions as negatives or rely on TF-IDF-based sampling), (2) the relative ranking of positive and negative samples, and (3) top-k selection (i.e., the number of negatives to include). Detailed results can be found in Table 2. Ultimately, selecting negatives from the model's incorrect predictions proved to be the most important factor, with an average score difference of 0.8%, while the other factors showed smaller differences of 0.5%. More importantly, regardless of the specific negative-aware training approach, its application leads to significant performance improvements compared to models like GenBioEL, which do not incorporate such training. All models applying negative-aware training, including our ANGEL model, outperformed GenBioEL by 0.9% to 1.7% ($p \leq 1.7e^{-5}$ for all comparisons).

Effect of similarity models Similarity models play a critical role in retrieving synonymous terms from the KB, and their choice can have a substantial impact on overall system performance. To assess their effectiveness, we evaluated three models: (1) TF-IDF, (2) BioBERT-NLI,[3] a sentence embedding model fine-tuned on natural language inference datasets, and (3) SapBERT. These models were incorporated into the positive-aware training framework of GenBioEL on the NCBI-Disease and BC5CDR datasets. As presented in Table 3, the TF-IDF-based approach outperforms the two embedding-based models. Although the strong performance of trigram-based similarity highlights the utility of surface-level matching in BioEL, this does not imply that the task is inherently simple. While many synonyms exhibit similar surface forms, a substantial portion do not—posing challenging edge cases that demand more nuanced semantic understanding.

[3] https://huggingface.co/gsarti/biobert-nli

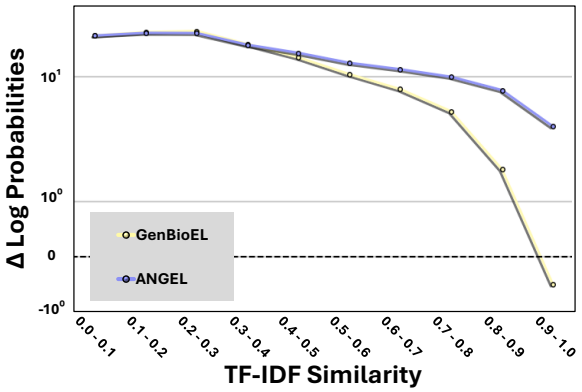

Figure 3: Analysis of the effect of negative-aware training. The x-axis represents the TF-IDF similarity between the input mentions and negative entities, while the y-axis depicts the difference in log probabilities between the top-1 positive prediction and negative entities for a given input mention. The NCBI-disease dataset was used.

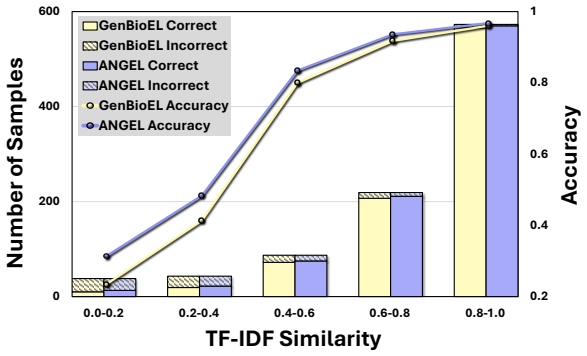

Figure 4: In-depth evalution of GenBioEL and our ANGEL models based on the TF-IDF similarity between the input mentions and gold-standard entities. The NCBI-disease dataset was used.

Effect of pre-training Table 4 highlights the effectiveness of ANGEL's pre-training by comparing other pre-training methods. BART, pre-trained using a standard language modeling objective but not specifically tailored for BioEL tasks, shows the lowest performance. In contrast, GenBioEL, pre-trained using synonyms from UMLS in a similar manner to our positive-only training, initially demonstrates a substantial performance advantage over BART. However, this gap narrows considerably after fine-tuning, to the point where it is no longer statistically significant. When ANGEL's negative-aware training is applied to GenBioEL, its performance improves significantly, achieving gains of 16.6% on BC5CDR and 10.9% on AAP. Even after fine-tuning, the performance gap remains noticeable, with a difference of 1.4% on BC5CDR and 0.9% on AAP.

5.2 Understanding the Effectiveness of Negative-aware Training

Figure 3 provide an interpretation of how negative-aware training leads to performance improvements. While positive-only training increases the probability of identifying synonyms, it also raises the risk of incorrectly boosting the probability of negative samples that are morphologically similar to the input mention. In contrast, negative-aware training improves the identification of synonyms while simultaneously reducing the probability of incorrect negatives, making it particularly effective when these negatives share morphological similarity with the input mention. To verify this, we divided input mention-negative pairs from the NCBI-disease dataset into 10 bins based on their tri-gram TF-IDF similarity. We then computed the log probabilities of negatives within each bin for the corresponding input mentions, comparing them to the log probabilities of the top-1 positive predictions (i.e., the synonym assigned the highest probability by the model). As the similarity between the input mention and the negative entities increased, the probabilities assigned by GenBioEL to positive and negative samples became more similar, eventually leading to higher probabilities for the negative samples. In contrast, our model demonstrated a clear distinction in behavior, consistently prioritizing positive samples over negatives.

5.3 Error Analysis

We conducted an in-depth evaluation of the models based on the similarity between the input mentions and the gold-standard entities. Similarity was calculated using tri-gram TF-IDF, with the gold-standard entity determined as the candidate synonym with the highest similarity score to the input mention. The similarity scores, ranging from 0 to 1.0, were divided into five bins, and accuracy was measured for each bin. As shown in Figure 4, errors predominantly occurred in the 0-0.2 and 0.2–0.4 bins, as indicated by the height of the hatched bars, which represent the number of errors. This suggests that models tend to struggle when the surface forms of the input mentions are not closely aligned with those of the gold-standard entities. Our method improves the generalizability of the model, leading to an overall reduction in GenBioEL's errors across all bins, with particularly notable improvements in

Rank	SapBERT	GenBioEL	ANGEL (Ours)
	... aggressive the same way someone with [ST] ***ASPD*** [ET] would be, except teenagers ...		(SNOMED CT:26665006)
1	ASP	Anankastic personality disorder	**Antisocial personality disorder (disorder)***
2	Acquired immune deficiency syndrome (disorder)	Borderline personality disorder	**Antisocial personality disorder***
3	Acquired immune deficiency syndrome	Oppositional defiant disorder	Borderline personality disorder (disorder)
4	Mesalazine	**Antisocial personality disorder***	Obsessive compulsive disorder (disorder)
5	Cryopyrin associated periodic syndrome (disorder)	Oppositional defiant disorder (disorder)	**Dissocial personality disorder***
	... I switched from lantus to [ST] ***basaglar*** [ET] in january and ...		(SNOMED CT:411529005)
1	Beagle	Linagliptin substance	**Insulin glargine substance***
2	Basiliximab sodium	Benzodiazepine substance	**Insulin glargine***
3	Basiliximab substance	Carisoprodol substance	Insulin glulisine substance
4	Albiglutide	Cariprazine	Ulipristal substance
5	Albiglutide substance	Benzocaine containing product	Lansoprazole
	... effects on amino acid (r-aminobutyric acid (GABA), [ST] ***glutamine*** [ET], aspartate and glutathione) levels ...		(MeSH:D018698)
1	Glutamine	Glutamine	L-glutamine
2	**Glutamic acid***	**Glutamic acid***	Glutamine
3	L-glutamine	Glutamylmethionine	D-glutamine
4	**L-glutamic acid***	Glutamylalanine	Glutamic acids
5	Glutamic acids	Glutaminic acids	**Glutamic acid***

Figure 5: Top-5 predictions from different BioEL models are presented. Entity names with correct identifiers are highlighted in boldface with an asterisk. The first and second examples highlight the strengths of our model, while the final example illustrates its limitations. For a detailed explanation, please refer to the main text.

cases of low similarity. However, significant challenges remain, as the accuracy of our model is only 34.2% in the 0–0.2 bin, highlighting the need for further improvement.

5.4 Case Study

Figure 5 illustrates the predictions of SapBERT, GenBioEL, and ANGEL. In the first example, the mention 'ASPD' is an abbreviation for 'antisocial personality disorder' (also known as 'dissocial personality disorder'). SapBERT incorrectly predicts 'ASP' due to the similarity in surface form. GenBioEL struggles to distinguish between correct entity names and those containing the words 'personality disorder'. In contrast, our model successfully identifies the correct entities, without being misled by false entity names that contain overlapping terms. The second example involves the mention 'basaglar,' a medication that contains insulin glargine, a long-acting insulin. The challenge here arises from the fact that product names can differ significantly from the biomedical terms used to describe their active ingredients. This discrepancy leads to failures in both SapBERT and GenBioEL, as they struggle to connect the brand name to its corresponding biomedical entity. Nevertheless, our model successfully identifies the correct entity, showcasing its ability to handle such complex cases effectively. In the final example, our method was less effective. For the mention of 'glutamine,' neither SapBERT nor GenBioEL identified the correct answer, but they did rank 'Glutamic acid,' the correct entity, within the top 5 candidates. Our model, however, ranked the correct answer slightly lower. Consequently, while our model shows a notable improvement in top-1 accuracy, the increase in top-5 accuracy is relatively modest in some datasets. The effectiveness of our method also varies across different datasets. We discuss this limitation in more detail in Appendix E, noting that such cases are an area for further exploration.

6 Conclusions

In this study, we discussed the importance of negative samples in training generative BioEL models and introduced ANGEL, the first framework in this field to effectively incorporate negative-aware training into a generative model. Our models demonstrated the ability to learn subtle distinctions between entities with similar surface forms and contexts. Experimental results showed that ANGEL outperformed existing similarity-based and generative models, with notable performance improvements of 0.9%, 1.4%, and 1.7% for BART, BioBART, and GenBioEL, respectively, achieving the best performance across five public BioEL datasets.

Limitations

Our method is versatile and applicable to any generative model, but it has only been tested on encoder-decoder models and not on decoder-only models such as BioGPT (Luo et al., 2022). We plan to further investigate the effect of our method on these models. Additionally, it has not been tested on recent open-source large language models (LLMs) (Touvron et al., 2023; Chen et al., 2023). While we acknowledge that incorporating comparisons with LLMs and further assessing the effectiveness of our approach would be an interesting direction, using LLMs for entity linking presents new challenges. The primary concern with larger models is their inefficiency, particularly regarding slower inference speeds and higher memory requirements, which may render them unsuitable for most real-world applications. This issue becomes particularly problematic in biomedical information extraction, where processing millions of publications to extract meaningful insights is essential.

Our negative-aware training method may not be limited to a specific domain, yet we have only evaluated it on biomedical-domain datasets, which restricts the demonstration of its broad applicability. Nevertheless, we would like to emphasize the reasons for focusing on the biomedical domain. Biomedical entity linking has unique characteristics that differentiate it from other domains, making this problem both challenging and interesting. In general domains, ambiguity typically arises between different types of entities (e.g., whether "Liverpool" refers to a city or a sports club). Similarly, in the biomedical domain, ambiguity exists between different types, such as whether "Ebola" in Figure 1 refers to a disease or a virus. Additionally, biomedical entities often exhibit significant variations in their surface forms, even when they share the same identifier, i.e., they refer to the same entity. As shown in Figure 5, "Basaglar" can be expressed as other variations such as "insulin glargine substance" or "insulin glargine." Furthermore, terms like "*substance*" in the entity "insulin glargine substance" overlap with many other entities (e.g., "Basiliximab *substance*," "Linagliptin *substance*," "Benzodiazepine *substance*"), making the task even more complex. Therefore, distinguishing between numerous candidates with similar surface forms is especially crucial in biomedical entity linking. We believe that our method, which trains the model using negative samples with similar structures, is particularly well-suited to tackle this challenge. However, exploring the application of our approach in other domains would be a valuable direction for future research.

Ethical Considerations

This study complies with ethical standards, ensuring that all datasets and models adhere to their respective licenses and usage terms. While our method was evaluated on five widely used datasets, these serve primarily as benchmarks and may not fully capture real-world complexities. Although the model demonstrates significant improvements, its limitations in handling low-similarity cases highlight the need for thorough validation before deployment, particularly in sensitive applications.

Acknowledgements

This research was supported by (1) the National Research Foundation of Korea (NRF-2023R1A2C3004176, RS-2023-00262002), (2) the Ministry of Health & Welfare, Republic of Korea (HR20C002103), (3) ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201819). M.S. was supported by (1) No. RS-2022-00155911: Artificial Intelligence Convergence Innovation Human Resources Development(Kyung Hee University), (2) No. RS-2024-00509257: Global AI Frontier Lab), and (3) the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2024- 00438239, 15%).

References

Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew Mccallum. 2022. Entity linking via explicit mention-mention coreference modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4644–4658.

Rico Angell, Nicholas Monath, Sunil Mohan, Nishant Yadav, and Andrew Mccallum. 2021. Clustering-based inference for biomedical entity linking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2598–2608.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137. Association for Computational Linguistics.

Rajarshi Bhowmik, Karl Stratos, and Gerard De Melo. 2021. Fast and effective biomedical entity linking using a dual encoder. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 28–37.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581):81.

Eunsuk Chang and Javed Mostafa. 2021. The use of snomed ct, 2013-2020: a literature review. *Journal of the American Medical Informatics Association*, 28(9):2017–2026.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Aaron M Cohen and William R Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71.

Allan Peter Davis, Thomas C Wiegers, Michael C Rosenstein, and Carolyn J Mattingly. 2012. Medic: a practical disease vocabulary used at the comparative toxicogenomics database. *Database*, 2012:bar065.

N De Cao, G Izacard, S Riedel, and F Petroni. 2020. Autoregressive entity retrieval. In *ICLR 2021-9th International Conference on Learning Representations*, volume 2021. ICLR.

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.

Ada Hamosh, Alan F. Scott, Joanna S. Amberger, Carol A. Bocchini, and Victor A. McKusick. 2004. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33:D514 – D517.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546.

Tuan Lai, Heng Ji, and ChengXiang Zhai. 2021. Bert might be overkill: A tiny but effective biomedical entity linker based on residual convolutional neural networks. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1631–1639.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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

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

Nut Limsopatham and Nigel Collier. 2016a. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1014–1023.

Nut Limsopatham and Nigel Collier. 2016b. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1014–1023.

Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multi-modal contrastive representation learning for entity alignment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2572–2584.

Zhenxi Lin, Ziheng Zhang, Xian Wu, and Yefeng Zheng. 2024. Biomedical entity linking as multiple choice question answering. In *International Conference on Language Resources and Evaluation*.

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

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Zulfat Miftahutdinov, Elena Tutubalina, Samsung-PDMI Joint AI Center, and PDMI RAS. 2019. Deep neural models for medical concept normalization in user-generated texts. *ACL 2019*, page 393.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Xiaoming Shi, Sendong Zhao, Yuxuan Wang, Xi Chen, Ziheng Zhang, Yefeng Zheng, and Wanxiang Che. 2021. Understanding patient query with weak supervision from doctor response. *IEEE Journal of Biomedical and Health Informatics*, 26(6):2770–2777.

Sunghwan Sohn, Donald C Comeau, Won Kim, and W John Wilbur. 2008. Abbreviation definition identification based on automatic precision estimates. *BMC bioinformatics*, 9:1–10.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Shogo Ujiie, Hayate Iso, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2021. End-to-end biomedical entity linking with span-based dictionary matching. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 162–167.

Yuejia Xiang, Ziheng Zhang, Jiaoyan Chen, Xi Chen, Zhenxi Lin, and Yefeng Zheng. 2021. Ontoea: Ontology-guided entity alignment via joint knowledge graph embedding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1117–1128.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: pushing the boundaries of llm performance in machine translation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 55204–55224.

Zhenran Xu, Yulin Chen, and Baotian Hu. 2023. Improving biomedical entity linking with cross-entity interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13869–13877.

Hongyi Yuan and Sheng Yu. 2024. Efficient symptom inquiring and diagnosis via adaptive alignment of reinforcement learning and classification. *Artificial Intelligence in Medicine*, 148:102748.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022a. Biobart: Pretraining and evaluation of a biomedical generative language model. *BioNLP 2022@ ACL 2022*, page 97.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022b. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4038–4048.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880.

Yongqi Zhang, Quanming Yao, Ling Yue, Xian Wu, Ziheng Zhang, Zhenxi Lin, and Yefeng Zheng. 2023. Emerging drug interaction prediction enabled by a flow-based graph neural network with biomedical network. *Nature Computational Science*, 3(12):1023–1033.

Baohang Zhou, Xiangrui Cai, Ying Zhang, and Xiaojie Yuan. 2021. An end-to-end progressive multi-task learning framework for medical named entity recognition and normalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6214–6224.

A Datasets

Table A presents the statistics of the five datasets used, along with their corresponding target knowledge bases.

NCBI-disease The NCBI-disease dataset (Doğan et al., 2014) contains 793 PubMed abstracts annotated with 6,892 disease mentions that are mapped to 790 unique disease concepts using the MEDIC ontology (Davis et al., 2012). MEDIC is a medical dictionary that integrates disease concepts, synonyms, and definitions from both MeSH (Lipscomb, 2000) and OMIM (Hamosh et al., 2004), encompassing a total of 9,700 unique disease entities. This dataset is primarily used for disease recognition and concept normalization tasks.

BC5CDR The BC5CDR dataset (Li et al., 2016b) includes 1,500 PubMed abstracts with 4,409 chemical entities, 5,818 disease entities, and 3,116 chemical-disease interactions. All annotated entities are mapped to the MeSH ontology (Lipscomb, 2000), which is a subset of UMLS (Bodenreider, 2004). This dataset is widely used for biomedical entity recognition and interaction studies. To fit the purpose of our study, we use only the chemical and disease annotations and discard the interaction annotations.

COMETA COMETA (Basaldella et al., 2020) focuses on layman medical terminology, compiled from four years of content across 68 health-related subreddits. This dataset consists of 20K biomedical entity mentions annotated with concepts from SNOMED CT (Chang and Mostafa, 2021). It is utilized for the normalization of consumer health expressions into standardized terminologies.

AskAPatient (AAP) The AskAPatient dataset (Limsopatham and Collier, 2016b) contains 8,662 phrases from social media language, each mapped to medical concepts from SNOMED CT (Chang and Mostafa, 2021). This dataset does not include contextual information, meaning that mentions are disambiguated solely based on the phrases themselves. Since the AskAPatient dataset lacks a test set, we employed a 10-fold cross-validation approach as outlined in the original paper by Limsopatham and Collier (2016a). The statistics reported are the averages across these folds.

MM-ST21pv The Medmentions dataset (Mohan and Li, 2019) is a large-scale resource for biomedical entity recognition. The ST21pv subset includes 4,392 PubMed abstracts with over 200,000 entity mentions linked to 21 selected UMLS semantic types. This dataset provides a comprehensive resource for training and evaluating biomedical entity recognition systems. Unlike the original dataset, we use the 2020AA version of UMLS as the KBs because the 2017AA version of UMLS is not directly accessible. This leads to some differences after preprocessing due to variations between versions. Specifically, our dataset deviates from the original Medmentions dataset by 741 training samples (0.6%), 284 validation samples (0.7%), and 235 test samples (0.6%).

B Hyperparameter Configurations

Table B details the hyperparameters used for positive-only training and negative-aware training across the BioEL benchmark datasets. We searched for the optimal hyperparameter settings using the validation sets. We refer to the study of Yuan et al. (2022b) to determine the range of the hyperparameters. During pre-training, we used the same hyperparameters as in GenBioEL. For positive-only training, we explored a range of training steps between 20K and 40K, a learning rate between 2e-5 and 3e-7, and batch sizes from 8 to 16, except during pre-training. During negative-aware training, we fixed the β at 0.1, in accordance with the basic configuration of DPO, and searched the hyperparameter space using a learning rate between 2e-5 and 1e-6 and batch sizes ranging from 8 to 64. We used the source codes provided by Yuan et al. (2022b)[4] and alignment handbook (Tunstall et al., 2023)[5].

C Ablation Study

Effect of pre-training In addition to Table 4 which shows the effect of pre-training on BC5CDR and AAP, Table C demonstrates that ANGEL's pre-training improves top-1 accuracy on NCBI-disease, COMETA, and MM-ST21pv, both before fine-tuning (✗) and after fine-tuning (✓).

Effect of optimization functions Our negative-aware framework is compatible with various optimization methods. To demonstrate this flexibility, we evaluated three additional loss functions during fine-tuning: (i) a simple pairwise loss, where

[4] https://github.com/Yuanhy1997/GenBioEL

[5] https://github.com/huggingface/alignment-handbook

Dataset	NCBI	BC5CDR	COMETA	AAP	MM-ST21pv
Entity types	Disease	Disease/chemical	Medical concepts	Medical concepts	21 UMLS types
# Examples					
Training	5,784	9,285	13,489	15,665	121,498
Validation	787	9,515	2,176	793	40,600
Test	960	9,654	4,350	866	39,922
KB statistics					
Entity names	108,092	809,929	904,798	3,398	6,051,091
Identifiers	14,944	268,162	350,830	1,036	3,092,324

Table A: The statistics of the benchmark datasets and their corresponding KBs.

Hyperparameter	Pre-training	Fine-tuning				
		NCBI	BC5CDR	COMETA	AAP	MM-ST21pv
Positive-only Training						
Training Steps	80K	20K	30K	40K	30K	40K
Learning Rate	4e-5	3e-7	5e-6	2e-5	5e-6	3e-5
Weight Decay	0.01	0.01	0.01	0.01	0.01	0.01
Batch Size	384	16	16	16	16	16
Adam ϵ	1e-8	1e-8	1e-8	1e-8	1e-8	1e-8
Adam β	(0.9,0.999)	(0.9,0.999)	(0.9,0.999)	(0.9,0.999)	(0.9,0.999)	(0.9,0.999)
Warmup Steps	1,600	0	500	1000	0	1,000
Attention Dropout	0.1	0.1	0.1	0.1	0.1	0.1
Clipping Grad	0.1	0.1	0.1	0.1	0.1	0.1
Label Smoothing	0.1	0.1	0.1	0.1	0.1	0.1
Negative-aware Training						
Epochs	5	1	1	1	1	1
Learning Rate	1e-5	1e-5	1e-6	5e-6	5e-6	5e-6
β (DPO)	0.1	0.1	0.1	0.1	0.1	0.1
Weight Decay	0.01	0.01	0.01	0.01	0.01	0.01
Batch Size	64	16	16	32	8	16
Warmup Steps	1000	-	-	-	-	-

Table B: Hyperparameters for positive-only training and negative-aware training.

Model	FT	NCBI	COMETA	MM
BART	✗	10.7	8.4	0.9
GenBioEL	✗	58.2	42.4	10.4
+ ANGEL (**Ours**)	✗	**64.6**	**49.8**	**18.2**
BART	✓	90.3	80.4	70.1
GenBioEL	✓	91.0	80.9	70.7
+ ANGEL (**Ours**)	✓	**92.8**	**82.8**	**73.3**

Table C: The top-1 accuracy of models with different pre-training strategies, along with the fine-tuned scores. 'FT' denotes fine-tuning, with ✗ representing pre-trained models without fine-tuning, and ✓ indicating models fine-tuned on human-annotated training sets. 'MM' represents the MM-ST21pv dataset.

Model	Acc@1
GenBioEL	85.0
ANGEL_{FT} (Pairwise)	85.9
ANGEL_{FT} (CPO) (Xu et al., 2024)	85.9
ANGEL_{FT} (SimPO) (Meng et al., 2024)	86.1
ANGEL_{FT} (DPO) (Rafailov et al., 2024)	86.2

Table D: Performance with different optimization functions. Average top-1 accuracy across the five benchmarks is reported.

$r_\theta(\mathbf{e} \mid \mathbf{x})$ is defined as $\log p_\theta(\mathbf{e}$; and two preference optimization methods that build upon and improve DPO: (ii) Contrastive Preference Optimization (CPO) (Xu et al., 2024), and (iii) Similarity Preference Optimization (SimPO) (Meng et al., 2024). Note that Equation 4 in the Method section presents a simplified version for clarity of explanation. In practice, CPO and SimPO introduce additional terms and require slight extensions to this formulation. For a detailed comparison of the exact equations, we refer readers to Table 7 in Meng et al. (2024), which provides a clear summary of the differences.

As shown in Table D, while DPO achieves the highest accuracy, the performance differences among the optimizers are relatively small. Regard-

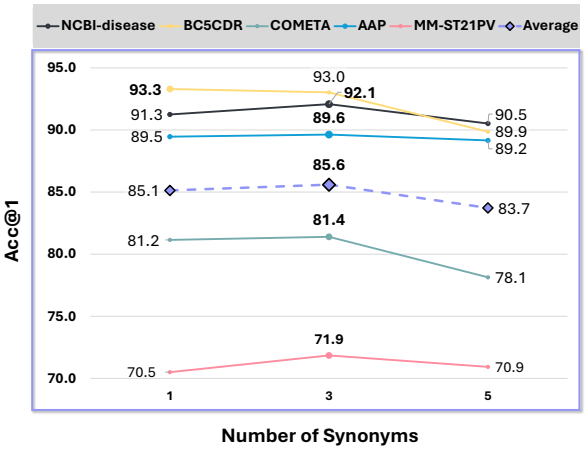

Figure A: The ablation study to determine the optimal number of synonyms. GenBioEL with ANGEL_{PT} was fine-tuned in this experiment. The scores are generally the highest when $k = 3$.

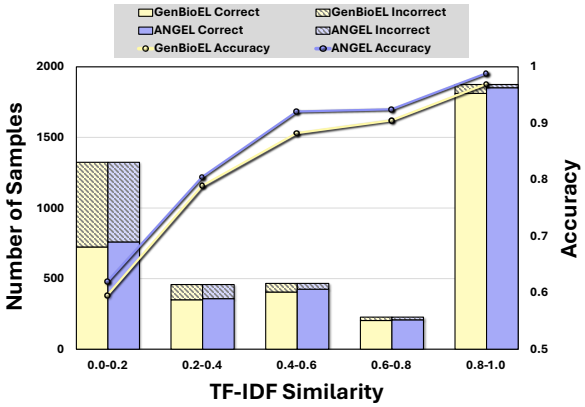

Figure B: In-depth evaluation of GenBioEL and ANGEL using TF-IDF similarity between input mentions and gold entities on the COMETA dataset.

less of the specific method used, all optimizers consistently enhance performance and outperform the GenBioEL baseline.

The number of synonyms To evaluate the impact of incorporating multiple synonyms during fine-tuning (Equation 3), we conducted experiments by varying the number of synonyms associated with each mention, testing with 1, 3, and 5 synonyms. As a result, using 3 synonyms proved to be optimal, outperforming the approach that used only a single top-1 synonym in the study of Yuan et al. (2022b).

D Error Analysis

Consistent with the analysis on the NCBI-disease dataset (Figure 4), Figure B reveals that models on the COMETA dataset most frequently made errors

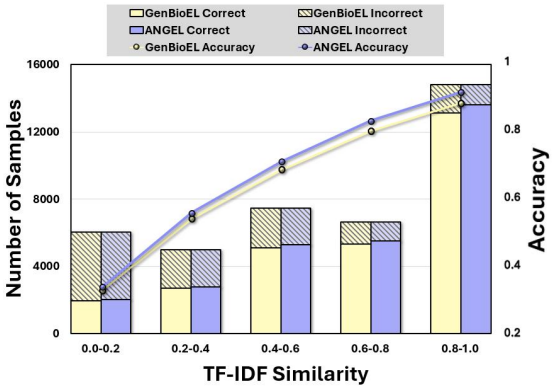

Figure C: In-depth evaluation of GenBioEL and ANGEL using TF-IDF similarity between input mentions and gold entities on the Medmentions dataset.

Model	BC5CDR		AAP	
	Acc@1	Acc@5	Acc@1	Acc@5
GenBioEL	93.1	95.7	89.3	**95.4**
+ ANGEL_{FT}	94.4	96.5	89.5	94.7
+ $\text{ANGEL}_{\text{PT + FT}}$	**94.5**	**96.8**	**90.2**	95.2

Table E: Comparison of top-1 and top-5 accuracy between the baseline model and models trained with ANGEL method after fine-tuning and pre-training on the BC5CDR and AAP datasets.

in the 0.0–0.2 bin, where input mentions have low similarity to gold-standard entities. Across all similarity bins, our ANGEL framework consistently improved upon GenBioEL's performance, leading to overall gains. A similar trend is observed in Figure C for the MedMentions dataset, where ANGEL again outperforms GenBioEL across all bins. These results highlight the need for future research focused on reducing errors in low-similarity scenarios.

E Top-5 Accuracy

Table E presents our model's top-1 and top-5 accuracy on the BC5CDR and AAP datasets. It compares the performance of our model in its baseline form (GenBioEL) and after fine-tuning (ANGEL_{FT}) and combined pre-training and fine-tuning ($\text{ANGEL}_{\text{PT + FT}}$). Our approach consistently boosts top-1 accuracy across all datasets, though the trends in top-5 accuracy are less uniform. In BC5CDR, both top-1 and top-5 accuracy show significant improvements: top-1 accuracy rises by 1.4 percentage points (from 93.1% to 94.5%), and top-5 accuracy increases by 1.1 percentage points (from 95.7% to 96.8%). However,

the AAP dataset exhibits a different pattern. While top-1 accuracy improves by 0.9 percentage points (from 89.3% to 90.2%), top-5 accuracy slightly declines: there is a 0.7 percentage points drop (from 95.4% to 94.7%) after fine-tuning and a 0.2 percentage points decrease (from 95.4% to 95.2%) after combined pre-training and fine-tuning. This decline in top-5 accuracy may be due to the AAP dataset's limited contextual information, forcing the model to rely predominantly on the mention form, making it more challenging to maintain high accuracy across multiple predictions. Additionally, the negative sampling strategy could unintentionally bias the model toward optimizing top-1 accuracy, thereby impacting top-5 performance. In conclusion, while our method consistently improves top-1 accuracy, the occasional slight decreases in top-5 accuracy, as observed in the AAP dataset, underscore the need for further refinement to maintain balanced accuracy across different ranking levels. Future work should focus on training strategies that preserve or enhance top-5 accuracy alongside top-1 improvements.