BEL HD : Improving Biomedical Entity Linking with Homonoym Disambiguation

Samuele Garda and **Ulf Leser**
Humboldt-Universität zu Berlin
{gardasam,leser}@informatik.hu-berlin.de

Abstract

Biomedical entity linking (BEL) is the task of grounding entity mentions to a knowledge base (KB). A popular approach to the task are name-based methods, i.e. those identifying the most appropriate name in the KB for a given mention, either via dense retrieval or autoregressive modeling. However, as these methods directly return KB names, they cannot cope with homonyms, i.e. different KB entities sharing the exact same name. This significantly affects their performance, especially for KBs where homonyms account for a large amount of entity mentions (e.g. UMLS and NCBI Gene). We therefore present **BELHD** (**B**iomedical **E**ntity **L**inking with **H**omonym **D**isambiguation), a new name-based method that copes with this challenge. Specifically, BELHD builds upon the BioSyn (Sung et al., 2020) model introducing two crucial extensions. First, it performs a preprocessing of the KB in which it expands homonyms with an automatically chosen disambiguating string, thus enforcing unique linking decisions. Second, we introduce *candidate sharing*, a novel strategy to select candidates for contrastive learning that enhances the overall training signal. Experiments with ten corpora and five entity types show that BELHD improves upon state-of-the-art approaches, achieving the best results in six out ten corpora with an average improvement of 4.55pp recall@1. Furthermore, the KB preprocessing is orthogonal to the core prdiction model and thus can also improve other methods, which we exemplify for GenBioEL (Yuan et al., 2022), a generative name-based BEL approach. Code is available at: link added upon publication.

1 Introduction

Biomedical entity linking (BEL) is the task of grounding text mentions to a Knowledge Base (KB) Approaches to BEL can be divided into two categories[1]. Entity-based methods explicitly construct

entity representations usually in form of embeddings for dense retrieval (Varma et al., 2021; Zhang et al., 2022; Agarwal et al., 2022 inter alia). In contrast, name-based methods directly identify the best matching name in the KB, either via dense retrieval or autoregressive modeling (see Figure 1).

Though name-based methods have been widely investigated and often outperform other approaches in evaluations (Sung et al., 2020; Liu et al., 2021; Yuan et al., 2022), they all suffer from a critical flaw, which is the treatment of homonyms. A homonym is a name in a KB that appears more than once, which results from multiple KB entities having the same name. As shown in Figure 1, since name-based methods return the name as a result of linking, they are not capable of resolving such cases (Zhang et al., 2022). Previous evaluations applied evaluation schemes in which the model is allowed to return multiple KB entities which are then counted as partly correct. However, in standard applications BEL is just one step in a complex pipeline in which downstream components typically cannot properly handle non-unique linking information (Wang et al., 2022).

We therefore introduce **BELHD** (**B**iomedical **E**ntity **L**inking with **H**omonym **D**isambiguation), which extends the name-based method BioSyn (Sung et al., 2020) to properly handle homonyms. That is, we disambiguate all KB names in a preprocessing step by expanding any instance of a homonym with a disambiguating string. An example is shown in Figure 1 (c), in which "Discharge" is a homonym referring the UMLS entities C0030685 and C0600083. We replace both instances with properly expanded names, i.e. "Discharge (Patient Discharge)" and "Discharge (Body Fluid Discharge)". Secondly, similar to (De Cao et al., 2021a), BELHD processes the entire text with *all* of its mentions at once. This allows us to introduce *candidate sharing*, a novel strategy for contrastive learning (Le-Khac et al., 2020) in

---

[1]We exclude methods requiring a candidate list, i.e. reranking approaches like cross-encoders (Wu et al., 2020).
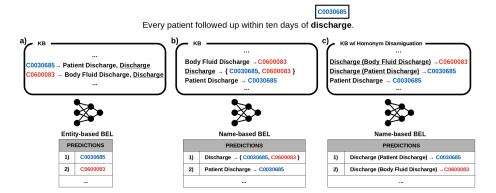
Figure 1: Illustration of entity-based (a) and name-based (b) approaches to biomedical entity linking. Underlined text highlights the KB homonym (Section 3) preventing a unique linking decision (b). In (c) we show how in BELHD we address the issue by replacing homonyms their disambiguated version. Text in blue and red represent the correct and wrong prediction, respectively.

which each mention not only uses its own candidates, but also those retrieved for other mentions appearing in the same context. As co-occurring mentions are often related, shared candidates can act as additional hard positive/negative samples enhancing the overall training signal (see Figure 3 for an example).

To evaluate our BELHD, we performed extensive experiments to compare its results to entity-based models, arboEL (Agarwal et al., 2022) and KRISSBERT (Zhang et al., 2022), and name-based ones, BioSyn (without our extensions) (Sung et al., 2020) and GenBioEL (Yuan et al., 2022), on ten corpora linked to six KBs. Overall, we find that BELHD outperforms state-of-the-art approaches in six of the ten corpora with an average improvement of 4.55pp recall@1. As our approach for homonym disambiguation is independent of the core model it can also be applied to improve other name-based methods, which we show for the case of GenBioEL.

2 Related Work

BEL models (excluding reranking ones) can be roughly divided into two categories (Section 1). The first one includes entity-based methods, i.e. those learning a representation for each entity. Agarwal et al. (2022) propose arboEL, the current state-of-the-art approach for MedMentions, which concatenates all entity names to form an entity embedding and constructs k-nearest neighbor graphs over co-referent mention and entity clusters. Using a pruning algorithm they generate directed minimum spanning trees rooted at entity nodes used for linking. Though not strictly entity-based,

KRISSBERT (Zhang et al., 2022) uses "entity prototypes" for linking, which are contextualized mentions of UMLS names (discarding homonyms) retrieved via exact string-matching from PubMed (one of the largest archive of biomedical literature). UMLS entities never mentioned in PubMed are represented by embeddings obtained with their names, semantic hierarchy and descriptions. The second category consists of name-based models, primarily bi-encoder architectures, all of them encoding mentions without context. Notable examples are (i) BioSyn (Sung et al., 2020), proposing a loss function to encourage KB names of the same entity to be closer in the dense space, (ii) SapBERT (Liu et al., 2021), presenting a pre-training strategy based on self-supervision shown to improve BioSyn performance and (ii) follow-up studies replacing BERT (Devlin et al., 2019) with a CNN to reduce the demanding memory footprint (Lai et al., 2021; Chen et al., 2021). Yuan et al. (2022) instead introduce GenBioEL, an adaptation for the biomedical domain of GENRE, the autoregressive approach to entity linking first introduced by De Cao et al. (2021b).

Homonyms are a well known issue in BEL (Wei and Kao, 2011; Leaman et al., 2013). Recently, Garda et al. (2023) introduce a comprehensive BEL benchmark and note that previous studies have focused on only two class of homonyms. The first one is abbreviations, e.g."TS" being either "Tourette Syndrome" or "Timothy Syndrome" while the second one are cross-species genes, e.g. human vs mouse "α2microglobulin". State-of-the-art approaches address these instances via specialized tools: Ab3P (Sohn et al., 2008) for abbreviations

and SR4GN (Wei et al., 2012) or SpeciesAssignment (Luo et al., 2022) for cross-species genes. These however are non-adaptable solutions crafted for a specific category of homonyms, leaving many cases unhandled, as the one in the example in Figure 1.

KB augmentation has been explored in previous work, though not specific for homonyms. Procopio et al. (2023) extend Wikipedia entities, represented by article titles, with their description in WikiData, e.g. "Ronaldo" with "Brazilian football player". To improve GENRE's generalization abilities Schumacher et al. (2023) propose to augment Wikipedia entities with keywords extracted from their descriptions with an unsupervised method. Both approaches however are solutions specific for Wikipedia, while ours targets the biomedical domain. Secondly, in BEL, entity descriptions are seldom available (Zhang et al., 2022).

3   Background

Task We formulate BEL as the task of predicting an entity $e \in E$ from a KB given a document d and a pair of start and end positions $\langle m_s, m_e \rangle$ indicating a span in d (the entity mention m). In all experiments we use in-KB (Röder et al., 2018) gold mentions. That is, after the BEL-step, each m is associated to a KB entity.

Biomedical KBs and homonyms. Each entity e in the KB is represented with a unique ID associated with a set of names $s \in S$. E.g. in UMLS the entity C0030685 has the following names: "Patient Discharge" and "Discharge". As shown in Figure 1, this entails that there are two ways to represent the KB: (a) by entity or (b) by name. The latter requires defining a mapping $\mathcal{V}_{KB} : S \to E$, which for a given s returns its associated entity. The exact same name however can appear more than once in the KB and thus points to multiple entities. In this case we call the name a *homonym*, "Discharge" in our example. Formally, a name s is a homonym if $|\mathcal{V}_{KB}(s)| > 1$, where $|\mathcal{V}_{KB}(s)|$ is the number of entities s maps to.

Impact of homonyms In case of name-based systems, homonyms fundamentally disrupt linking, as for mentions linked to them it is impossible to assign a unique KB entity. Determining to what extent this issue impacts name-based methods in general is however non-trivial. This is because the number of mentions linked to homonyms depends on the specific model. For instance, for the mention "discharge" in Figure 1, it is possible for a name-based system to return "Patient Discharge". However, due to the high similarity between surface forms, we expect most current models to rank "Discharge" higher. We can obtain an *approximate* estimate by considering mentions affected by homonyms if their gold KB entity has an associated name (a) which is a homonym and (b) whose surface form is highly similar to the mention's one. Concretely, we consider a mention to be highly similar to a KB name if their normalized Levenshtein distance (Marzal and Vidal, 1993) is one (see Appendix A for details).

In Table 1 we report this estimate for corpora in BELB (Garda et al., 2023), a BEL benchmark comprising ten commonly used BEL corpora linked to six KBs (see Appendix B for details). Though some corpora have no such cases, we see that despite being only ~2% of UMLS names (see Table 2 for exact counts), homonyms may account for up to 26% of the links in the widely used MedMentions, imposing an upper bound to the performance of name-based methods. The amount rises steeply to more than 60% for genes, for which homonyms are a well known aspect (Wei and Kao, 2011) as the same gene can be found in multiple species (see §4.1.1).

4   Method

We now introduce (i) our approach to replace KB homonyms with a disambiguated version and (ii)

BELB: Biomedical Entity Linking Benchmark (Garda et al., 2023)		
KB (entity type)		Homonyms
Corpus	Affected mentions	
CTD DISEASES (Disease)		0.39%
NCBI Disease	1.25% (12 / 960)	
BC5CDR (D)	0.18% (8 / 4,363)	
CTD CHEMICALS (Chemical)		≪1%
BC5CDR (C)	0% (0 / 5,334)	
NLM-Chem	0% (0 / 11,514)	
CELLOSAURUS (Cell line)		3.21%
BioID	3.47% (30 / 864)	
NCBI GENE (Gene)		53.61%
GNormPlus	68.84% (2,218 / 3,222)	
NLM-Gene	66.1% (1,804 / 2,729)	
NCBI TAXONOMY (Species)		0.04%
Linnaeus	0% (0 / 1,430)	
S800	0% (0 /767)	
UMLS		2.07%
MedMentions (st21pv)	26.41% (10,602 / 40,143)	

Table 1: Relative number of homonyms and approximate estimate of the links they account for in BELB (test set).
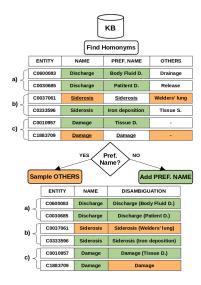
Figure 2: Illustration of our Homonym Disambiguation approach for biomedical KBs.

our architectural enhancements to BioSyn, which together result in BELHD, our novel method for BEL.

4.1 Homonym Disambiguation

In BELHD we modify how the KB is represented to address the homonym issue. As mentioned, a homonym is a KB name s s.t. $|\mathcal{V}_{\mathcal{KB}}(s)| = n$ with $n > 1$, where n is the number of associated entities. For our approach we draw inspiration from Wikipedia, where article titles that would otherwise be homonymous are disambiguated via additional information in parentheses[2]. Similarly, we expand every homonym into into n different versions, one for each entity, each augmented with a string having distinguishing information on the entity it represents. For instance we replace "Discharge" with "Discharge (Patient Discharge)" and "Discharge (Body Fluid Discharge)". The augmentation strings are other names in the KB. The procedure used to select them assumes that the KB reports for each entity, which of its associated names is the *preferred* one (usually the official name).

Our approach proceeds as follows (see Algorithm 1 for pseudocode). As shown in Figure 2, we first collect all names s in the KB which are homonyms (see Appendix C for details) and other names associated to the entities they refer to. If the homonym is not the preferred name, we create disambiguated versions by adding the preferred name of the entities they represent, as in (a). If in-

stead the homonym is itself the preferred name, we select as disambiguation string the *shortest* name (which is not the homonym) associated to the entity, as in (b). This simple strategy can be replaced with a KB-specific solution if metadata is available, e.g. selecting the name's official long form. Finally, there are cases in which extending the name is not possible. This happens for homonyms which are preferred names but not do provide additional names, as in (c). However, if a homonym has n associated entities, we only need to create $n - 1$ disambiguated versions, with the unmodified one acting as the default meaning.

4.1.1 Cross-species homonyms

The approach described above is valid for any biomedical KB whose entities specify a preferred name. However, in BEL there exists a special class of homonyms which requires an additional step for complete disambiguation. Specific to Gene and Cell line, these are cross-species homonyms. For instance, in NCBI GENE "α2microglobulin" can refer both to the human and cattle gene. As both genes have "A2M" as preferred name, our approach would generate two still identical entries "α2microglobulin (A2M)"[3].

Cross-species homonyms play such a crucial role for these entity types that the corresponding KBs (NCBI GENE and CELLOSAURUS) always report for each entity the species as well, in form of NCBI TAXONOMY entities. E.g. NCBI GENE for the human "A2M" reports 9606 ("human") while for the cattle one 9913 ("cattle"). Therefore, before applying our first procedure (resolving intra-species homonyms), we identify all cross-species homonyms (see Appendix C for details) and generate disambiguated versions with the species name from NCBI TAXONOMY. If a name is both an intra- and cross-species homonym, it will have both disambiguation strings. E.g. "A2M" is also a secondary name for the human gene "IGHA2", so our approach will generate both "A2M (α2microglobulin, human)" and "A2M (IGHA2, human)".

4.2 BELHD

Here we first review the model upon which our approach is built, i.e. BioSyn, whose key aspect is its objective function. Sung et al. (2020) observe

According to ICD-11 a diagnosis of [S] Tourette's Syndrome [E] cannot follow one of [S] vocal tic disorder [E].
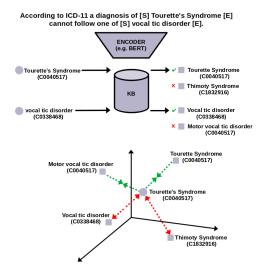
Figure 3: Overview of a BELHD training step with *candidate sharing*.

that having a name-based search space entails that there are potentially *multiple* valid candidates for m. Therefore, they propose to train BioSyn as follows. For a given mention m, BioSyn uses dense retrieval (maximum inner product search) to fetch a set of candidates name $C = \{c_i, \cdots c_n\}$ from the (pre-encoded) KB[4]. The probability of each c_i to be a correct link for m is defined as:

$$\mathrm{P}(c_i|m) = \frac{\exp(sim(m, c_i))}{\sum_{j=1}^{|C|} \exp(sim(m, c_j))} \quad (1)$$

where $|C|$ is the size of C and sim is the inner product $\langle \cdot, \cdot \rangle$. The model is optimized to minimize the following *marginal maximum likelihood* (MML):

$$l_m = -\log \sum_{i=1}^{|C|} \mathbb{1}_{[\mathcal{V}_\mathcal{C}(m)=\mathcal{V}_{\mathcal{KB}}(c_i)]} \mathrm{P}(c_i|m) \quad (2)$$

where $\mathcal{V}_\mathcal{C} : M \to E$ is a mapping returning the associated KB entity for a given mention $m \in M$, while $\mathbb{1}_{[i=j]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $i = j$. Intuitively speaking the objective function encourages the representations of m and all candidates names to be close in the dense space if they are associated the same KB entity.

To improve BioSyn, we keep the same training objective and introduce the following changes (see §5.2.2 for ablation study).

Retrieval For simplicity, we described the BioSyn variant using only dense retrieval. The original

model combines both sparse (bi-gram TF-IDF) and dense retrieval to fetch candidates from the KB. For BELHD, we rely exclusively on the latter as the impact of sparse candidates in BioSyn is minimal.

Context In contrast to BioSyn, BELHD leverages contextual information. For this we mark mentions boundaries with two special tokens: $[S]$ and $[E]$. To obtain a single embedding we use mean pooling over $[S]$ and $[E]$.

Projection head To reduce memory footprint and increase search speed, we introduce a projection head parametrized by a weight matrix $W \in \mathbb{R}^{h \times p}$ applied to *both candidates and mentions* embeddings, where h and p are the original and the projection head size, respectively.

Candidate sharing As reported by Gillick et al. (2019), hard negative mining is crucial to successfully training (B)EL approaches based on contrastive learning (like BioSyn). This entails retrieving difficult candidates for a mention from the KB at fixed intervals (e.g. at the end of each epoch). We note however that, as co-occurring mentions are often related (Chen et al., 2021), an important and easily exploitable training signal are candidates retrieved for other mentions in the same document. For instance in Figure 3, we see that by sharing candidates across mentions it is possible to explicitly increase the similarity between "Tourette's Syndrome" and "Motor vocal tic disorder", which is another valid name for Tourette's, but unlikely to be retrieved during hard negative mining due to its completely different surface form (even for models not based on lexical matching).

Therefore, similar to De Cao et al. (2021a), we encode the entire input text with all of its mentions. However, while they use a Longformer (Beltagy et al., 2020) to support long sequences, we keep a BERT-based architecture[5]. To overcome the maximum sequence length we split each text into sentences and treat them as a single mini-batch. Then for each mention $m_i \in M = \{m_i \cdots m_n\}$ we create pool of candidates C_i. Half of the candidates are retrieved directly from the KB (as in BioSyn). The other half are instead selected from $\bigcup_{\substack{j=1 \\ j \neq i}}^{|M|} C_j$ (candidates of other mentions different from those of m_i), choosing the ones most similar to m_i (measured by sim). Intuitively, if $\mathbb{1}_{[\mathcal{V}_\mathcal{C}(m_i)=\mathcal{V}_{\mathcal{KB}}(c_j)]} = 1$, these are hard positives (as

---

[4]For simplicity, with a slight abuse of notation, we use m and c_i to refer both to the strings and their embeddings.

[5]To the best of our knowledge there is no Longformer pretrained on biomedical data. The clinical version provided by Li et al. (2023) performed poorly on preliminary experiments.

the example in Figure 3), hard negatives otherwise.

5 Experiments

In this section we perform empirical studies to (i) validate the effectiveness of our HD procedure and (ii) compare our novel method against the state-of-the-art.

5.1 Setting

Corpora and KBs In all experiments we rely on BELB (Garda et al., 2023) to access corpora and KBs. As a standardized benchmark, BELB removes confounding factors such as differences in preprocessing and KB versions. BELB consists of ten corpora linked to six KBs covering five entity types, which allows for an evaluation more comprehensive than the one reported in previous studies. In Table 1 we report BELB corpora and KBs used in our experiments (see Appendix B for details). As shown in Garda et al. (2023), current implementations of BEL systems cannot scale to large KBs like NCBI GENE (>40M/100M entities/names). Therefore, for NCBI GENE, we use the subsets determined by the species of the genes in `GNormPlus` and `NLM-Gene` (see Appendix F). This reflects a common real-world use case, since often only a specific subset of species is relevant for linking (e.g. human and mouse).

Training Unless otherwise stated, we ensure that all models receive the same training signal by (i) training them from scratch on BELB corpora and (ii) avoiding corpus- or KB-specific pre-training. For all baselines we rely on the original implementation provided by the authors. Details w.r.t. models and training can be found in Appendix E.

Metric We report mention-level micro-average recall@1. If for a given mention a method returns multiple entities as prediction, instead of resorting to random sampling as proposed by Zhang et al. (2022), we consider the prediction incorrect. This is because a primary use case of (B)EL is its deployment into pipelines (Wang et al., 2022), where multiple entities cannot be used.

5.2 Homonym Disambiguation

The effectiveness of our approach in disambiguating homonyms is measured by its success rate, i.e. the ratio between homonyms which after expansion are no longer homonyms and the original ones in the KB. From Table 2 we see that our approach can disambiguate virtually all homonyms in all six KBs

(see Appendix F for NCBI GENE subsets). The failure cases are KB names having the same surface form *after* our approach extends them with distinguishing information, and are not to be confused with the case where a preferred name is a homonym but has no alternative names, as case (c) in §4.1. We find that these are caused by a combination of two factors. The first is related to the quality of the KBs. UMLS and NCBI GENE are meta-KBs, i.e. they integrate entities from multiple KBs. In few cases, this causes them to have two distinct entities having however little to no difference in terms of names. The second is using the shortest alternative name as disambiguation string when a homonym is also the entity's preferred name. E.g. in UMLS C0003663 and C0020316 have almost identical list of associated names. C0003663's preferred name is "Aquacobalamin", having as secondary name "Hydroxocobalamin". However, "Hydroxocobalamin" is also C0020316's preferred name, whose shortest alternative name is "Aquacobalamin", giving raise to two "Hydroxocobalamin (Aquacobalamin)".

5.2.1 Results

From Table 3 we see that BELHD is the overall best approach, outperforming all baselines in six out of ten corpora, with GenBioEL as second. Our results are in line with general-domain entity linking, where name-based approaches generally outperform those relying on entity representations (Schumacher et al., 2023). Importantly, we note that our HD solution can be used with any name-based model. Notably, when equipped with HD, GenBioEL's performance increases by 63.54pp on the homonym-rich `NLM-Gene`, outperforming all other methods, including BELHD. This can be attributed to the fact that the corpus was specifically created to test models on cross-species homonyms (Islamaj et al., 2021). Key to success on the task is contextual information, primarily in form of species mentions. While GenBioEL processes the entire text, BELHD uses sentences (see §4.2), thus being unable to access species information if it does not occur in the same sentence as the gene mention. Finally, we see that HD as minimal to no effect in BioSyn, since it does not use context. BioSyn is however the best model for the `Linnaues` corpus. This can be explained by the fact that `Linnaeus` was created specifically for the development of dictionary-based approaches (Luoma et al., 2023), giving a strong advantage to methods using string-matching like BioSyn.

	CTD DISEASES (Disease)	CTD CHEMICALS (Chemical)	CELLOSAURUS (Cell line)	NCBI GENE (Gene)	NCBI TAXONOMY (Species)	UMLS
Entities	13,188	175,663	144,568	42,252,923	2,491,364	3,464,809
Names	88,548	451,410	251,747	105,570,090	3,783,882	7,938,833
Homonyms	349 (0.39%)	2 ($\ll$1%)	8,070 (3.21%)	56,597,279 (53.61%)	1,422 (0.04%)	164,154 (2.07%)
- pref. name	1	-	502	822,103	27	12,530
- other	348	2	2,582	20,965,362	1,395	151,624
- cross-species	-	-	5,416	56,265,683	-	-
Success rate	100%	100%	100%	$\gg$99% (1054)	100%	$\gg$99% (39)

Table 2: Number of names and homonyms (% relative to names) categorized by cases (see §4.1) in BELB KBs. The success rate is the ratio between homonyms which after augmentation are no longer homonyms and the original amount. Number of failures, i.e. names that are still homonyms (duplicates) is reported in parenthesis.

	CTD DISEASES (Disease)		CTD CHEMICALS (Chemical)		CELLOSAURUS (Cell line)	NCBI GENE (Gene)		NCBI TAXONOMY (Species)		UMLS
	NCBI Disease	BC5CDR (D)	BC5CDR (C)	NLM-Chem	BioID	GNormPlus	NLM-Gene	S800	Linnaeus	MedMentions
Entity-based										
arboEL †	80.00	84.87	87.40	71.76	95.02	34.64	29.96	78.62	74.97	<u>68.67</u>
KRISSBERT † ‡	82.80	85.0	**95.10**	-	-	-	-	-	-	61.30
Name-based										
BioSyn	79.90	84.83	84.57	70.35	80.79	OOM	OOM	82.79	88.60	OOM
+ HD	79.90	$84.23_{-0.60}$	$85.00_{+0.43}$	$71.31_{+0.96}$	$81.60_{+0.9}$	OOM	OOM	$81.23_{-1.56}$	$\mathbf{88.81}_{+0.21}$	OOM
GenBioEL	82.71	<u>88.29</u>	<u>94.60</u>	<u>75.00</u>	94.79	6.80	2.89	88.27	76.92	41.16
+ HD	$\underline{83.02}_{+0.31}$	$88.20_{-0.09}$	$94.15_{-0.45}$	$74.10_{-0.90}$	$\underline{96.30}_{+1.51}$	$\underline{66.08}_{+59.28}$	$\underline{66.43}_{+63.54}$	$\mathbf{89.96}_{+1.69}$	$77.62_{-0.30}$	$\underline{64.59}_{+23.43}$
BELHD (ours)	**87.60**	**89.23**	92.93	**82.39**	**96.99**	**77.84**	<u>59.03</u>	<u>84.35</u>	<u>81.89</u>	**70.58**

Table 3: Performance of all models on BELB corpora (test set). **Bold** and <u>underlined</u> indicate best and second best score, respectively. HD: Homonym Disambiguation (§4.1). OOM: out-of-memory (>200GB) † Without cross-encoder reranking ‡ Authors provide code only for the "supervised" variant[6], i.e. entity prototypes are based on train/development mentions. As this variant cannot link zero-shot entities we report results from Table 7 in (Zhang et al., 2022) for a fair comparison.

5.2.2 Ablation study

	NCBI GENE (Gene)
	NLM-Gene
BELHD (ours)	59.03
- HD	$6.67_{-52.8}$
- context	$32.72_{-26.31}$
- candidate sharing	$56.83_{-2.20}$
- projection head	$58.74_{-0.29}$

Table 4: Ablation study of improvements over BioSyn introduced in BELHD (see §4.2).

Table 4 reports our ablation study of BELHD improvements over BioSyn (see §4.2). We see that the most important component is HD, which is critical for a homonym-rich entity type like Gene, while the use of contextual information is second. The third strongest improvement is brought by *candidate sharing*. This confirms that leveraging the relatedness of co-occurring mentions enhances training signal improving overall results. Finally, the projection head, despite reducing the embedding

dimensionality, does not hurt performance.

5.2.3 Ad-hoc solutions for Homonym Disambiguation

	UMLS	NCBI GENE (Gene)	
	MedMentions	GNormPlus	NLM-Gene
arboEL	68.67	34.64	29.96
+ AR	$68.85_{+0.18}$	$35.66_{+1.02}$	$30.30_{+0.34}$
+ Species †	-	$\mathbf{47.83}_{+13.19}$	$\mathbf{40.82}_{+10.86}$
GenBioEL	41.16	6.80	2.89
+ AR	$42.01_{+0.85}$	$7.67_{+0.87}$	$3.48_{+0.59}$
+ SA	-	$65.70_{+58.90}$	$61.78_{+58.89}$
+ HD	$\mathbf{64.59}_{+23.43}$	$66.08_{+59.28}$	$66.43_{+63.54}$
BELHD w/o HD	57.23	13.90	6.67
+ AR	$59.50_{+2.27}$	$15.98_{+2.08}$	$6.85_{+0.18}$
+ SA	-	$43.67_{+29.77}$	$42.58_{+35.91}$
+ HD	$\mathbf{70.58}_{+13.35}$	$\mathbf{77.84}_{+63.94}$	$\mathbf{59.03}_{+52.35}$

Table 5: Effect of different strategies to handle homonyms. HD: Homonym Disambiguation (ours), AR: Abbreviation Resolution (Sohn et al., 2008), SA: Species Assignment (Luo et al., 2022). † Include species name into entity representation (Kartchner et al., 2023)

Here we compare ad-hoc methods to handle homonyms (see Section 2) with our general HD approach on the corpora most affected by homonyms.

---

[6]https://huggingface.co/microsoft/BiomedNLP-KRISSBERT-PubMed-UMLS-EL

We perform abbreviation resolution (AR) with Ab3P (replacing abbreviations with their long form) and retrain all models. Secondly, we identify and assign species to gene mentions with SpeciesAssignment (SA) and filter predictions accordingly. As AR resolves difficult mentions which are not necessarily affected by homonyms (see Section 3) we include the entity-based arboEL in the comparison. As SA is specific to name-based methods, for arboEL we include species name in entity representations as proposed by (Kartchner et al., 2023). From Table 5 we see that HD delivers the best results across corpora. It significantly outperforms AR, confirming that addressing a wider range of homonyms is critical. Interestingly, HD outperforms the highly specialized SA approach as well. We argue that this is due to species information not being always explicitly expressed (Wei et al., 2012), upon which SA relies. HD instead is more versatile, allowing the model to *learn* useful contextual patterns beyond explicit species mentions.

5.2.4 Entity- vs name-based dense retrieval

	CTD DISEASES (Disease)	
		NCBI Disease
Entity-based		72.19
Name-based (w/ HD)		78.85

Table 6: Difference between entity-based and name-based (with HD) approach with same experimental conditions (input representation, model, candidate selection).

Large corpora like `MedMentions` are the exception in BEL (see Appendix B). We therefore aim to compare the name-based and entity-based approach when limited training data is available. To exclude potential confounding factors (input representation, model type, candidate selection) we follow the experimental setting of Wu et al. (2020) and train two identical bi-encoders[7] on the same input (see Appendix E.1 for details). The models differ only in the KB representation and, consequently, in the objective function: one uses standard cross entropy (entity-based) while the other MML (name-based: see §4.2). Results in Table 6 suggest that name-based are inherently more sample-efficient,

as they can directly leverage surface similarities between mentions and KB names, while entity-based require more training data to optimize entity representations.

6 Discussion

We introduce BELHD, a novel method for biomedical entity linking. Our experiments show that BELHD outperforms all baselines in six out of ten BEL corpora. We stress that we retrain all models on BELB with the code and hyperparameters provided by the authors and thus numbers we report may differ from those found in the original publications (see Section 8). However, we believe that this setting is the best approximation to fairly compare across methods since, as reported by Garda et al. (2023), BEL studies present stark differences in preprocessing and experimental setups making comparison based on published numbers problematic.

Secondly, we note that our study focuses on *first-stage* BEL, i.e. candidate generator methods. Reranking, e.g. with a cross-encoder (Wu et al., 2020), is a further enhancement *orthogonal* to the choice of the generator. Thus using a cross-encoder only for one generator (arboEL) as in (Kartchner et al., 2023) gives it an unfair advantage producing biased results. Either different rerankers are compared with the same set of candidates or the same reranker assesses the quality of different sets. We leave an analysis of the latter as future work.

7 Conclusion

We highlight how homonyms in biomedical KBs significantly impact performance of BEL methods returning KB names as predictions. We introduce BELHD, a novel BEL approach based on BioSyn (Sung et al., 2020) outperforming all baselines in six out of ten corpora. We show that its primary feature HD is a general solution improving results in other name-based methods as well.

8 Limitations

A limitation of our work is the assumption on the biomedical KBs required by the HD procedure. That is, the KB must specify which is the preferred name of a given entity. Though, to the best of our knowledge, virtually all biomedical KBs meet this assumption, the HD procedure is therefore not strictly KB-agnostic. Secondly, due to quality issues in biomedical KBs, i.e. entities having almost

---

[7]We focus on dense retrieval since De Cao et al. (2021b) already shown their generative approach performs poorly if trained to generate unique IDs instead of names.

identical lists of associated names, it is not possible to remove all homonyms (see §4.1). This could be mitigated by using more sophisticated strategies to select the disambiguation string, e.g. using the least similar name determined by Levenshtein distance, which we leave as future work .

As already mentioned in Section 6, an important limitation of our study is the lack of hyperparameter exploration of all baselines. Due to the high computational resources necessary to train BEL models we are limited to rely on the default ones reported by the authors. It is therefore possible that optimizing them may result in better numbers. Additionally, due to the large amount of combinations of corpora, models and ad-hoc components we are limited to our best effort and do not report results with multiple seeds. We note as well that we train all models from scratch on BELB corpora avoiding corpus- or KB-specific pre-trained weights. This is because, as noted by Milich and Akbik (2023), different methods use different pre-training strategies and different data, ultimately impairing direct comparison. It is therefore possible that GenBioEL, with its *KB-Guided Pre-training*, and BioSyn, by using SapBERT weights (Liu et al., 2021), may achieve higher results. Finally, we note that there exists entity-specific BEL models, e.g. GNorm2 (Wei et al., 2023) for genes. We do not evaluate them as the primary focus of this study are entity-agnostic BEL models.

9 Ethical Considerations

A primary use case of biomedical entity linking is its deployment in information extraction pipelines, which in turn are deployed in application to facilitate the navigation of the scientific literature by biomedical researchers. As shown by an existing large body of work, language models (such as the one used in our work) may present biases. Therefore, a potential harmful downstream consequence may be casused by systems' errors caused by those biases. For instance, a system incorrectly linking a mention of "Postmenopausal Osteoporosis" (C0029458) to the general "Osteoporis" (C0029456) due to an implicit gender bias may prevent a relevant publication for the condition to be found by researchers. Secondly, if results of extraction pipelines are used to populate biomedical knowledge bases, and in turn these resources are used to train other models, these implicit biases may be further propogated and amplified.

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, Seattle, United States. Association for Computational Linguistics.

Cecilia Arighi, Lynette Hirschman, Thomas Lemberger, Samuel Bayer, Robin Liechti, Donald Comeau, and Cathy Wu. 2017. Bio-ID track overview. In *BioCreative VI Challenge Evaluation Workshop*, volume 482, page 376.

Amos Bairoch. 2018. The Cellosaurus, a cell-line knowledge resource. *Journal of biomolecular techniques: JBT*, 29(2):25.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

O. Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:267D–270.

Garth R. Brown, Vichet Hem, Kenneth S. Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D. Pruitt, Donna R. Maglott, and Terence D. Murphy. 2015. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Research*, 43(D1):D36–D42.

Lihu Chen, Gaël Varoquaux, and Fabian M Suchanek. 2021. A lightweight neural model for biomedical entity linking. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12657–12665.

Allan Peter Davis, Thomas C Wiegers, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, and Carolyn J Mattingly. 2023. Comparative Toxicogenomics Database (CTD): update 2023. *Nucleic Acids Research*, 51:D1257–D1262.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021a. Highly parallel autoregressive entity linking with discriminative correction. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021b. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Samuele Garda, Leon Weber-Genzel, Robert Martin, and Ulf Leser. 2023. BELB: a biomedical entity linking benchmark. *Bioinformatics*.

Martin Gerner, Goran Nenadic, and Casey M Bergman. 2010. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):1–17.

Dan Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Rezarta Islamaj, Robert Leaman, David Cissel, Cathleen Coss, Joseph Denicola, Carol Fisher, Rob Guzman, Preeti Gokal Kochar, Nicholas Miliaras, Zoe Punske, Keiko Sekiya, Dorothy Trinh, Deborah Whitman, Susan Schmidt, and Zhiyong Lu. 2022. NLM-Chem-BC7: manually annotated full-text resources for chemical entity annotation and indexing in biomedical articles. *Database*, 2022.

Rezarta Islamaj, Chih-Hsuan Wei, David Cissel, Nicholas Miliaras, Olga Printseva, Oleg Rodionov, Keiko Sekiya, Janice Ward, and Zhiyong Lu. 2021. NLM-Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. *Journal of biomedical informatics*, 118:103779.

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

David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie Mitchell. 2023. A comprehensive evaluation of biomedical entity linking models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14462–14478, Singapore. Association for Computational Linguistics.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29:2909–2917.

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(baw068).

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30:340–347.

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, Online. Association for Computational Linguistics.

Ling Luo, Chih-Hsuan Wei, Po-Ting Lai, Qingyu Chen, Rezarta Islamaj, and Zhiyong Lu. 2022. Assigning species information to corresponding genes by a sequence labeling framework. *Database*, 2022.

Jouni Luoma, Katerina Nastou, Tomoko Ohta, Harttu Toivonen, Evangelos Pafilis, Lars Juhl Jensen, and Sampo Pyysalo. 2023. S1000: a better taxonomic name corpus for biomedical information extraction. *Bioinformatics*, 39.

Andrés Marzal and Enrique Vidal. 1993. Computation of normalized edit distance and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15:926–932.

Marcel Milich and Alan Akbik. 2023. Zelda: A comprehensive benchmark for supervised entity disambiguation. In *Proceedings of the 17th Conference of*

the European Chapter of the Association for Computational Linguistics, pages 2061–2072, Dubrovnik, Croatia. Association for Computational Linguistics.

Sunil Mohan and Donghui Li. 2019. MedMentions: A large biomedical corpus annotated with umls concepts. In In Proceedings of the 2019 Conference on Automated Knowledge Base Construction (AKBC 2019).

Evangelos Pafilis, Sune P. Frankild, Lucia Fanini, Sarah Faulwetter, Christina Pavloudi, Aikaterini Vasileiadou, Christos Arvanitidis, and Lars Juhl Jensen. 2013. The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. PLOS ONE, 8(6):e65390.

Luigi Procopio, Simone Conia, Edoardo Barba, and Roberto Navigli. 2023. Entity disambiguation with entity definitions. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1297–1303, Dubrovnik, Croatia. Association for Computational Linguistics.

Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. 2018. GERBIL – benchmarking named entity recognition and linking consistently. Semantic Web, 9:605–625.

Elliot Schumacher, James Mayfield, and Mark Dredze. 2023. On the surprising effectiveness of name matching alone in autoregressive entity linking. In Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCHING 2023), pages 58–69, Toronto, ON, Canada. Association for Computational Linguistics.

Federhen Scott. 2012. The NCBI Taxonomy database. Nucleic Acids Research, 40:D136–D143.

Sunghwan Sohn, Donald C Comeau, Won Kim, and W John Wilbur. 2008. Abbreviation definition identification based on automatic precision estimates. BMC Bioinformatics, 9.

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, Online. Association for Computational Linguistics.

Maya Varma, Laurel Orr, Sen Wu, Megan Leszczynski, Xiao Ling, and Christopher Ré. 2021. Cross-domain data integration for named entity disambiguation in biomedical text. Findings of the Association for Computational Linguistics: EMNLP 2021.

Xing David Wang, Ulf Leser, and Leon Weber. 2022. Beeds: Large-scale biomedical event extraction using distant supervision and question answering. In Proceedings of the 21st Workshop on Biomedical Language Processing, pages 298–309, Dublin, Ireland. Association for Computational Linguistics.

Chih-Hsuan Wei and Hung-Yu Kao. 2011. Cross-species gene normalization by species inference. BMC Bioinformatics, 12.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2012. SR4GN: A species recognition software tool for gene normalization. PLOS ONE, 7:e38460.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2015. GNormPlus: An integrative approach for tagging genes, gene families, and protein domains. BioMed Research International, 2015:e918710.

Chih-Hsuan Wei, Ling Luo, Rezarta Islamaj, Po-Ting Lai, and Zhiyong Lu. 2023. GNorm2: an improved gene name recognition and normalization system. Bioinformatics, 39.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6397–6407, Online. Association for Computational Linguistics.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022. 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, Seattle, United States. Association for Computational Linguistics.

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.

A   Approximate String Matching

We first preprocess mentions and KB names by lowercasing and removing all non alphanumeric characters. We compute a score in $[0, 1]$ (the higher the more similar) between each mention and the names associated with its gold KB entity and select the KB names with a similarity score of 1. The score is the defined as $\frac{D}{|s_i|+|s_j|}$, where D is the Levenshtein distance (with a substitution weight of 2) between strings s_i and s_j and $|s_i|$ is the number of characters in s_i. We use the implementation provided by https://github.com/maxbachmann/RapidFuzz.

B   Biomedical Entity Linking Benchmark

BELB is a biomedical entity linking benchmark introduced by Garda et al. (2023). It provides access to 10 corpora linked to six knowledge bases

| Entity type | | | |
KB	Entities	Names	Avg. names per entity
Disease			
CTD DISEASES (Davis et al., 2023)	13,188	88,548	6.71
Chemical			
CTD CHEMICALS (Davis et al., 2023)	175,663	451,410	2.56
Cell line			
CELLOSAURUS (Bairoch, 2018)	144,568	251,747	1.74
Species			
NCBI TAXONOMY (Scott, 2012)	2,491,364	3,783,882	1.51
Gene			
NCBI GENE (Brown et al., 2015)	42,252,923	105,570,090	2.49
GNormPlus subset	703,858	2,455,772	3.48
NLM-Gene subset	873,015	2,913,456	3.33
UMLS			
UMLS (Bodenreider, 2004)	3,464,809	7,938,833	2.29

Table 7: Overview of the KBs available in BELB according to their entity type. We report the number of entities, names and average name per entities

| Entity type | | | |
Corpus	Documents (train / dev / test)	Mentions (train / dev / test)	0-shot mentions
Disease			
NCBI Disease (Doğan et al., 2014)	592 / 100 / 100	5,133 / 787 / 960	150 (15.62%)
BC5CDR (D) (Li et al., 2016)	500 / 500 / 500	4,149 / 4,228 / 4,363	388 (8.89%)
Chemical			
BC5CDR (C)(Li et al., 2016)	500 / 500 / 500	5,148 / 5,298 / 5,334	1,038 (19.46%)
NLM-Chem † (Islamaj et al., 2022)	80 / 20 / 50	20,796 / 5,234 / 11,514	3,908 (33.94%)
Cell line			
BioID ‡ (Arighi et al., 2017)	231 / 59 / 60	3,815 / 1,096 / 864	158 (18.29%)
Species			
Linnaeus † (Gerner et al., 2010)	47 / 17 / 31	2,115 / 705 / 1,430	385 (26.92%)
S800 (Pafilis et al., 2013)	437 / 63 / 125	2,557 / 384 / 767	363 (47.33%)
Gene			
GNormPlus (Wei et al., 2015)	279 / 137 / 254	3,015 / 1,203 / 3,222	2,822 (87.59%)
NLM-Gene (Islamaj et al., 2021)	400 / 50 / 100	11,263 / 1,371 / 2,729	1,215 (44.52%)
UMLS			
MedMentions (st21pv) (Mohan and Li, 2019)	2,635 / 878 / 879	122,178 / 40,864 / 40,143	8,167 (20.34%)

Table 8: Overview of the corpora available in BELB with number of documents, mentions and 0-shot mentions (mentions linked to an entity not in the train/development set). Pairing of corpora and KB is determined by the entity type. ‡ Full text ‡ Figure captions

and spanning five entity types: Gene, Disease, Chemical, Species and Cell lines: see Table 7 and Table 8 for an overview of the KBs and corpora, respectively. The key feature of BELB is its standardized preprocessing of corpora and KBs, offering tight integration between the two. This makes it a standardized testbed which removes confounding factors such as differences in preprocessing and KB versions. All corpora consists of biomedical publications in English. For a detailed description and license information of each corpus and KB we refer the reader to the original publication.

C   Identify homonyms

```sql
CREATE TABLE kb(
    uid INTEGER PRIMARY KEY,
    -- entity label
    identifier INTEGER NOT NULL,
    -- pref. name (0), abbr. (1), ...
    description INTEGER NOT NULL,
    name TEXT NOT NULL,
    -- NCBI Taxonomy entity
    species INTEGER DEFAULT NULL
)
```

Listing 1: Schema used in BELB to store biomedical KBs.

In Listing 1 we present the unified schema provided in BELB used to store all biomedical KB. Each name is associated to an identifier (entity label) and a description, i.e. whether it is e.g. the preferred name or the abbreviated form. In case of KBs having cross-species homonyms (see §4.1.1) BELB stores the name of the associated species entity coming from NCBI TAXONOMY.

```
-- Homonyms
SELECT name FROM kb
    GROUP BY name,species
    HAVING count(*)>1

-- Cross-species homonyms
SELECT name FROM kb
    GROUP BY name
    HAVING count(*)>1 AND
    COUNT(DISTINCT(species))>1;
```

Listing 2: Example SQL queries to compute set of homonyms with BELB KBs.

Listing 2 shows how using BELB we generate the set of homonyms and cross-species homonyms. In the group by of the first query we include "species" to ensure that the homonyms belong to the same species (intra-species), e.g. "BRI3" can be either NCBI GENE 81618 or 25798, both however are *human* genes. The second query instead specifically identifies only cross-species homonyms, i.e. it constraints names to have different associated species.

D Pseudocode for Homonym Disambiguation procedure

In Algorithm 1 we present the pseudocode of our approach to resolve homonyms in biomedical KBs present in §4.1. The pseudocode for the cross-species procedure described in §4.1.1 can be easily derived from it.

E Models and training details

Here we report training details for all models considered in our study. We stress that we retrain all models on BELB with the code provided by the original authors (see Table 9 for links to implementations). All experiments were performed on two NVIDIA A100 GPUs.

BioSyn uses the default hyper-parameters provided by the authors. Unlike in the original study, we

---

Algorithm 1 Pseudocode our disambiguation approach.

Require: $\mathcal{H}$ ▷ Pre-computed set of homonyms
Require: $\mathcal{E}$ ▷ Entities
 1: **for each** $e \in \mathcal{E}$ **do**
 2:      $\mathcal{S} \leftarrow$ get_names(e)
 3:      **for each** $s \in \mathcal{S}$ **do**
 4:          **if** $s \in \mathcal{H}$ **then**
 5:              $p \leftarrow$ get_preferred_name$(\mathcal{S})$
 6:              **if** $s = p$ **then**
 7:                  ▷ s.t. $s \mathrel{!}= p$
 8:                  $d \leftarrow$ get_longest$(s, \mathcal{S})$
 9:              **else**
10:                  $d \leftarrow p$
11:              **end if**
12:              $s \leftarrow$ concatenate(s, d)
13:          **end if**
14:      **end for**
15: **end for**

	Implementation (link)
arboEL	https://github.com/dhdhagar/arboEL
GenBioEl	https://github.com/Yuanhy1997/GenBioEL
BioSyn	https://github.com/dmis-lab/BioSyn

Table 9: Implementation links of the biomedical entity linking models used in our experiments.

exclusively train on the train split of corpora (no development). Total amount of parameters: 110M.

GenBioEl uses different values for learning rate and warmup steps for `NCBI Disease` and `BC5CDR`. We cannot perform a full hyper-parameter search for each corpus, and therefore select the values that work best for both corpora, i.e. a learning rate of $1e-5$ and 500 warmup steps. Total amount of parameters: 400M.

arboEL's inference procedure is parametrized by the number of k nearest neighbor used to construct the graph (determining which pairs of nodes are connected). The implementation provided by the authors runs the inference trying different $k \in \{0, 1, 2, 4, 8\}$. For fair comparison with other models, we do not perform any hyperparameter optimization and hence report the score for $k = 0$. Total amount of parameters: 110M.

BELHD keeps all BioSyn hyper-parameter besides (i) the number of training epochs which we increase from 10 to 20 and the number of candidates k for each mention m_i, which we set to 32: 16 mention-specific and 16 from candidate sharing (see §4.2) Like in Biosyn, the KB is re-encoded at the and of each training epoch to keep the name embeddings consistent with the updated model parameters (Guu et al., 2020). The only exception is `MedMentions`, for which, due to its large size, we use 10 epochs and re-encode the KB every 1000 steps. The dimensionality of the projection head is set to 128. In order to fit the entire text unit at once, we split it into sentences (with segtok[8]) and treat it as a single mini-batch, using on gradient accumulation to achieve a larger batch size, which we set to 8. We rely on FAISS (Johnson et al., 2019) for efficient exact maximum inner product search. Total amount of parameters: 110M (projection head has 1e4 parameters).

E.1 Name- vs entity-based dense retrieval

Following Wu et al. (2020) both models take as input a mention centered in a fixed context window. The input is truncated (left and/or right) to be of maximum 128 tokens. The input to the candidate encoder (entity or name) is maximum 128 tokens. Each model uses in-batch negatives with an additional 10 hard negatives mined during training, i.e the top 10 predicted entities/names for each training mention. Both models are trained with a mini batch size of 32 for 10 epochs.

---

[8] https://github.com/fnl/segtok

F Corpus-specific NCBI Gene subsets

	NCBI GENE (Gene)	
	GNormPlus	NLM-Gene
Names	2,455,772	2,913,456
Homonyms	1,163,255 (47.37%)	1,479,719 (50.79%)
- pref. name	33,919	35,032
- other	323,824	39,1290
- cross-species	1,094,531	1,410,006
Success rate	>99% (520)	>99% (523)

Table 10: Equivalent to Table 2 for NCBI GENE corpus-specific subsets.

NCBI TAXONOMY		
Entity	Name	Corpora
3055	Chlamydomonas reinhardtii	NLM-Gene
3702	thale cress	GNormPlus,NLM-Gene
3847	soybean	GNormPlus
4896	fission yeast	GNormPlus,NLM-Gene
6239	Caenorhabditis elegans	GNormPlus,NLM-Gene
6956	European house dust mite	NLM-Gene
7227	fruit fly <Drosophila melanogaster>	GNormPlus,NLM-Gene
7955	zebrafish	GNormPlus,NLM-Gene
8355	African clawed frog	GNormPlus,NLM-Gene
8364	tropical clawed frog	GNormPlus,NLM-Gene
9031	chicken	GNormPlus,NLM-Gene
9606	human	GNormPlus,NLM-Gene
9615	dog	NLM-Gene
9823	pig	GNormPlus,NLM-Gene
9913	cattle	GNormPlus,NLM-Gene
9940	sheep	NLM-Gene
9986	rabbit	GNormPlus,NLM-Gene
10029	Chinese hamster	NLM-Gene
10089	Ryukyu mouse	NLM-Gene
10090	house mouse	GNormPlus,NLM-Gene
10116	Norway rat	GNormPlus,NLM-Gene
10298	Herpes simplex virus type 1	GNormPlus
11676	Human immunodeficiency virus 1	GNormPlus,NLM-Gene
11709	Human immunodeficiency virus 2	NLM-Gene
11908	Human T-cell leukemia virus type I	GNormPlus
41856	Hepatitis C virus genotype 1	GNormPlus
51031	New World hookworm	NLM-Gene
81972	Arabidopsis lyrata subsp. lyrata	NLM-Gene
333760	Human papillomavirus type 16	GNormPlus
511145	Escherichia coli str. K-12 substr. MG1655	GNormPlus
559292	Saccharomyces cerevisiae S288C	GNormPlus,NLM-Gene
2886926	Escherichia phage P1	NLM-Gene

Table 11: NCBI GENE subsets determined by the species (NCBI TAXONOMY entities) of the gene mentions in GNormPlus and NLM-Gene

For the NCBI GENE subsets determined by the species of the genes in GNormPlus and NLM-Gene (see §5.1) we report in Table 10 the number of homonyms and the success rate of our disambiguation approach. In Table 11 we report the NCBI GENE subsets determined by the species (NCBI TAXONOMY entities) of the gene mentions in GNormPlus and NLM-Gene.