OSLAT: Open Set Label Attention Transformer for Medical Entity Retrieval and Span Extraction

Raymond Li* RAYMONDL@CS.UBC.CA
University of British Columbia

Ilya Valmianski ILYA@CURAI.COM
Curai

Li Deng† L.DENG@IEEE.ORG
Vatic Investments

Xavier Amatriain XAVIER@CURAI.COM
Curai

Anitha Kannan ANITHA@CURAI.COM
Curai

Abstract

Medical entity span extraction and linking are critical steps for many healthcare NLP tasks. Most existing entity extraction methods either have a fixed vocabulary of medical entities or require span annotations. In this paper, we propose a method for linking an open set of entities that does not require any span annotations. Our method, **Open Set Label Attention Transformer (OSLAT)**, uses the label-attention mechanism to learn candidate-entity contextualized text representations. We find that OSLAT can not only link entities but is also able to implicitly learn spans associated with entities. We evaluate OSLAT on two tasks: (1) span extraction trained without explicit span annotations, and (2) entity linking trained without span-level annotation. We test the generalizability of our method by training two separate models on two datasets with low entity overlap and comparing cross-dataset performance.

Keywords: Clinical NLP, Attention, Open Set, Disjoint Spans, Contrastive Learning

1. Introduction

Many natural language processing (NLP) tasks in the healthcare domain such as information retrieval (IR) (Tamine and Goeuriot, 2021), diagnosis coding (Crammer et al., 2007), and conversational agents (Compton et al., 2021; Valmianski et al., 2021) greatly benefit from correctly identifying medical entities such as disorders and findings in the text. This has led to a wealth of literature centered on entity recognition in the past decades (Fries et al., 2020; Friedman et al., 1995; Chapman et al., 2001; Aronson, 2001; Savova et al., 2010) and many competitions/ tasks in both NLP and IR communities (Pestian et al., 2007; Styler et al., 2014; Elhadad et al., 2015; Bethard et al., 2016).

However, the problem of entity recognition continues to be largely unsolved, with two main challenges being: (1) lack of sufficient amounts of labeled and diverse data and (2)

* work done as a research intern at Curai
† work done while at Curai

Entity	Text containing the entity
knee swelling	pain and swelling in knee
knee pain	pain and swelling in knee
cervical lymphadenopathy	swollen lymph node on right side of neck
dyspnea	head pressure and anxiety for the past couple weeks also, having to take really deep breaths to catch my breath .

Table 1: We can see that the entity can present as a contiguous-span of text (row 1), disjoint-spans (rows 2-3) or overlapping-spans (row 1-2). For each (text, entity), the color saturation highlights the prediction confidence from OSLAT.

the ability to handle previously unseen entities (open-set recognition).

Existing methods require significant amounts of labeled data (Esteva et al., 2019) and often include two parts: (1) entity span annotations, and (2) span-entity linking annotations. For formal clinical texts, such as medical literature and physician notes, weak-labeling approaches (*e.g.* lookup-based) help reduce the need for span annotations (Fries et al., 2020). However, these approaches struggle with patient-derived text due to insufficient vocabulary coverage and the propensity of patient text to have disjoint spans for entities (see Table 1).

The in-the-wild open-set recognition challenge appears when the models are exposed to text containing entities not seen during training (Prabhu et al., 2019; Mottaghi et al., 2020). Even when using UMLS (Bodenreider, 2004) as the basis for medical vocabulary, it is difficult to collect enough data to cover *all* entities. Furthermore, real-life text often contains medically relevant compositional entities (*e.g.* "severe sudden abdominal pain") and colloquial language not in UMLS.

In this paper, we tackle both challenges through **Open Set Label Attention Transformer (OSLAT)**. OSLAT is similar, and similarly computationally efficient, to a bi-encoder information retrieval architecture. However, unlike bi-encoders, it uses the label-attention mechanism to create candidate-entity contextualized document representations, which can then be used to classify the presence of the candidate entity. Like bi-encoders, OSLAT can link open set entities and does not require span annotations. However, the label attention mechanism allows it to implicitly learn to infer entity span masks, even for disjoint spans.

We summarize our work by outlining the two technical contributions:

1. We use a transformer-based encoder to encode not only the input text but also the candidate labels in a label-attention architecture. This allows us to operate on an open set of labels.

2. To train our model, we introduce Label Synonym Supervised Normalized Temperature-Scaled Cross-Entropy (LSS-NT-Xent) loss, an extension of NT-Xent (Chen et al., 2020)

We test the generalizability of our approach by performing extensive experiments on two

datasets. Including evaluating on unseen entities, and applying a model trained on one dataset to a test set in the other dataset.[1]

2. Definition and Tasks Formulations

We begin by defining a universe of all entities, denoted by $\mathcal{E}$. Note, we do not need to explicitly define $\mathcal{E}$. During training, we will observe a subset of these entities $\mathcal{E}_{seen}$ and the remaining unobserved (open-set) is $\mathcal{E}_{unseen} = \mathcal{E} \setminus \mathcal{E}_{seen}$. We then assume access to a dataset $\mathcal{D}_{train} = \{(\mathbf{x}_t, e_t)\}_{t=1}^{T}$, where $\mathbf{x}_t$ is the t^{th} target text and e_t is an entity present in it, with $\mathcal{E}_{seen} = \cup_{t=1}^{T} e_t$. Note that entity mentions/spans are not available during training. For each entity, $e_t \in \mathcal{E}_{seen}$, we also assume access to its synonyms, obtained from an external source such as UMLS (which we use in this paper).

Task 1: Entity Span Extraction. For entity span extraction, we are provided with input text-entity pair $(\mathbf{x}, e)$ s.t. $e \in \mathcal{E}$, this reflects the application of the model in the wild. The goal of this is to identify the spans of text in $\mathbf{x}$ that describe e.

Task 2: Entity Linking. For entity linking, we are provided with the input text and a finite universe of entities $\mathcal{E}_{test} \subseteq \mathcal{E}$. The goal of this task is to predict whether entity e is mentioned in text $\mathbf{x}$, denoted as $s(x, e)$ s.t. $e \in \mathcal{E}_{test}$, which we refer to as the retrieval score.

3. Approach

Our method, **Open Set Label Attention Transformer (OSLAT)**, consists of separately encoding the entity and the text using a single encoder with the representations combined using label attention. In particular, the mean pooled representation of the entity is used to construct a query vector while the token embeddings of the text are used to construct the key-value vector pairs. This query, keys, and values are then used in a softmax attention to compute a single vector (OSLAT representation). The spans are inferred by looking at the label attention scores over the text tokens. This architecture is trained in two stages (see Figure 1 for an overview). In the first stage, we perform self-alignment training of the encoder (§ 3.1). In the second stage, we use our extension of NT-Xent, which we call LSS-NT-Xent, to contrastively train the OSLAT representations (§ 3.2). Finally, we perform entity linking by training a binary classifier on top of the OSLAT representations (§ 3.3).

3.1. Self-Alignment Pretraining on Medical Entities

We use BioBERT (Lee et al., 2020) as the backbone encoder. We perform self-alignment pretraining to decrease representational anisotropy of entity embeddings (Li et al., 2020; Carlsson et al., 2021; Gao et al., 2021; Liu et al., 2021a,b) (for the change in anisotropy see Figure C.1 in Appendix C). In particular, for medical entity $e_i \in \mathcal{E}_{seen}$, we obtain its representation $h^{(e_i)}$ by taking the [CLS] token embedding of the last hidden layer of BioBERT. To apply the contrastive loss function, we follow the model architecture described in SimCLR (Chen et al., 2020), where a two-level feed-forward projection head maps the representation $h^{(e_i)}$ from BioBERT into a low-dimension space, before a supervised contrastive loss, NT-Xent, is applied to the normalized projection output z_i (Chen et al., 2020; Khosla et al., 2020; Gao

1. Our implementation publicly is available. at: https://github.com/curai/curai-research/tree/main/OSLAT

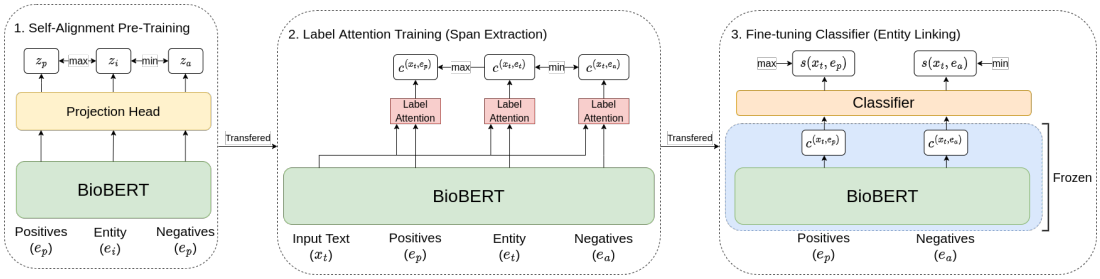


Figure 1: Overview of our proposed two-stage training approach. We first perform self-alignment pretraining on medical entities (§ 3.1), before aligning the label-text joint representations obtained through label attention (§ 3.2). Finally, we train a entity linking classifier on top of the frozen representations (§ 3.3).

et al., 2021):

$$\mathcal{L}_{\text{pre}} = \sum_{i \in B} \frac{-1}{|P(i)|} \left(\sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \sim \mathcal{E}_{seen}} \exp(z_i \cdot z_a / \tau)} \right) \quad (1)$$

For each entity in batch B, the positives z_p are projected representations from the synonym set $P(i)$ of entity e_i, with $|P(i)|$ as its cardinality, while the negatives z_a are projected representations from sampled entities from $\mathcal{E}_{seen}$. Finally, hyperparameter τ denotes the scalar temperature. As the entities are organized into disjoint synonym sets, we apply a stratified sampling strategy for sampling negatives, where we first sample a synonym set and then sample an entity from that set. This ensures that entities with a smaller synonym set do not get under-represented during training. After the self-alignment pretraining, we discard the projection head keeping the fine-tuned encoder. Details on our training procedure can be found in § 6.1.

3.2. Label Attention Training

OSLAT supports an open set of labels by jointly encoding labels and target texts into the same subspace. To obtain the representation of the entity spans within the target text, we first encode label e_t and target text $\mathbf{x}_t$ with our self-alignment pretrained BioBERT (see § 3.1). Specifically, for $(\mathbf{x}_t, e_t) \in \mathcal{D}$, the label representation $h^{(e_t)} \in \mathbb{R}^{1 \times d}$ and target text representation $h^{(x_t)} \in \mathbb{R}^{n \times d}$ from the last hidden layer of BioBERT (with hidden size d) are used to compute the label-attention score using a variant of the dot-product attention:

$$\alpha^{(\mathbf{x}_t, e_t)} = \text{Softmax}\big(h^{(e_t)} (h^{(x_t)})^T\big) \quad (2)$$

where the attention score $\alpha_k^{(\mathbf{x}_t, e_t)}$ can be interpreted as the token-wise semantic similarity between the label e_t and the kth token of target text $\mathbf{x}_t$. Since the [CLS] token for the target text can contain aggregate semantic information about the entire input, we found that the model often resorted to attending solely to the [CLS] token. To mitigate this issue, we remove the [CLS] token from $h^{(x_t)}$ to encourage the model to attend to other portions of the target text. Finally, we compute the entity span representation as a weighted sum of the target text $h^{(x_t)}$ by the attention scores:

$$c^{(\mathbf{x}_t, e_t)} = \sum_{k=1}^{n} \alpha_k^{(\mathbf{x}_t, e_t)} h_k^{(x_t)} \quad (3)$$

To train our model, we use a variant of NT-Xent which we call Label Synonym Supervised Normalized Temperature-Scaled Cross Entropy (LSS-NT-Xent):

$$\mathcal{L}_{\text{LSS}}(I) = \sum_{t \in B} \frac{-1}{|P(t)|} \left(\sum_{p \in P(t)} \log \frac{\exp(c^{(\mathbf{x}_t, e_t)} \cdot c^{(\mathbf{x}_t, e_p)} / \tau)}{\sum_{a \sim \mathcal{E}_{seen}} \exp(c^{(\mathbf{x}_t, e_t)} \cdot c^{(\mathbf{x}_t, e_a)} / \tau)} \right) \tag{4}$$

Similar to the self-alignment pre-training described in § 3.1, we use e_t's synonym set $P(t)$ as positives and randomly sample negatives from the $\mathcal{E}_{seen}$ and their synonyms. At inference time, we use the attention scores $\alpha^{(\mathbf{x}_t, e_t)}$ to predict whether each token of x_t lies in the span of entity e_t.

3.3. Entity Linking

For the task of entity linking, we apply a binary classifier on top of the entity span representation $c^{(x_t, e)}$ to predict the probability of entity e being mentioned in text x_t. During training, we optimize the classifier with Focal Loss (Lin et al., 2017), where the parameters of the OSLAT encoder are kept frozen to ensure that its attention weights can still be used for span extractions. For each example, the positives are synonyms of mentioned entities, while negatives are sampled from the universe of all entities. In practice, training time can be significantly reduced by caching the entity span representations of the training set in the first epoch.

At inference time, we use the classifier output as the unnormalized retrieval score:

$$s(x_t, e) = \text{Classifier}(c^{(x_t, e)}) \tag{5}$$

4. Related Work

Entity mention/Span detection. Unlike OSLAT, most approaches (*c.f.* Pagad and Pradeep (2022) for a comprehensive survey) for this task require access to a manually-labeled span-level dataset during training. Notable exceptions include Fries et al. (2020) that uses weak supervision to construct a labeled training set for the task or Fu et al. (2021) that learns to reconcile outputs from multiple span detection approaches. However, models used to generate inputs to Fu et al. (2021) require manual annotations.

Entity linking. Most methods (*c.f.* Sevgili et al. (2020) for a survey) require datasets with explicit spans during training, and often operate within a closed set of entities. While Wu et al. (2019); Mottaghi et al. (2020) operates with an open set, Wu et al. (2019) requires labeled entity spans while Mottaghi et al. (2020) sidesteps entity mention problem and poses it as a classification task within active learning paradigm. In contrast, OSLAT takes as input an entity of interest, and then simultaneously detects a span of text and whether the entity of interest is in that span.

Label attention in healthcare. Label attention for classification over a fixed label set is studied in various healthcare applications (Mullenbach et al., 2018; Vu et al., 2021; Liu et al., 2021c; Hu et al., 2018; Mayya et al., 2021; Nguyen and Ji, 2021).

Most previous approaches used it within a convolutional (Mullenbach et al., 2018; Liu et al., 2021c; Hu et al., 2018) or BiLSTM Vu et al. (2021) architecture for the task of classifying clinical notes into International Classification of Diseases (ICD) codes. Meanwhile, some recent works have also used transformer architecture (Mayya et al., 2021; Nguyen and Ji, 2021) for this task. In particular, Nguyen and Ji (2021) extends Mayya et al. (2021) by incorporating the label attention module into the encoder fine-tuning

process, leading to improvements in the encoder representations of biomedical text.

However, these previous approaches focus on the problem of multi-label classification over a fixed label set. In contrast, OSLAT can infer disjoint spans in the input text that map to an entity within an open set. To the best of our knowledge, this is the first work to address this problem in an open-set context.

5. Datasets

We are interested in investigating the following empirical questions:

- **Open set entity detection**:
  Is our approach robust to entities that are unseen during training?
- **Cross-domain transfer**:
  Is our approach robust when applied to data from a different domain than that of training (*e.g.* train on patient written and apply to provider/expert-written text)?
- **Handling disjoint-spans**:
  Is our approach robust in identifying entity spans that are disjoint?

In order to answer these questions, we build two complementary datasets. The first dataset (§ 5.2) is comprised of texts in which patients describe their health issues (RFE dataset). The second dataset (§ 5.3) is comprised of discharge summary notes written by physicians (hNLP dataset). The train-test split procedure (§ 5.1) of these datasets is itself non-trivial as we need to split both target texts and medical entities such that the test set contains both *seen* and *unseen* entities. Lastly, it's worth mentioning that there is a significant difference between the entity sets in both datasets (85% from hNLP to RFE and 69% from RFE to hNLP), this ensures that we do not provide undue advantage to the model during cross-domain evaluations. A detailed comparison between the two datasets is available in Appendix A.2.

5.1. Train/Test dataset construction

We start with an intermediate dataset of the form $(\mathbf{x}_k, E_k)$ where $\mathbf{x}_k$ is the k^{th} input text that has a set E_k of entities to reflect that multiple entities can be in the same input text. Then, $\mathcal{E} = \cup_k E_k$ is the universe of entities, and $p(e)$ is the marginal probability of entity e in the dataset.

Constructing $\mathcal{E}_{seen}$, $\mathcal{E}_{unseen}$: For our experiments, we choose 10% of the entities as *unseen*. We choose these entities randomly from 20%, 40%, and 40% from high, medium, and low marginal probability bins of $p(e)$ so that we capture entities across the spectrum of frequency distribution.

Train-Test split: We split the dataset into disjoint sets for training and testing from the perspective of the entity. For each entity $e \in \mathcal{E}_{unseen}$, we associate all pairs $((\mathbf{x}_k, e)_{k:e\in E_k})$ to the test set. For each entity $e \in \mathcal{E}_{seen}$, we randomly sample, without replacement, 10% of $\mathbf{x}_k, e)_{k:e\in E_k}$ pairs for the test set and remaining 90% to training set. We ensure that all entities in $\mathcal{E}_{seen}$ have at least five examples in the training set. If not, we first prioritize adding to the training set.

Span level labels for test set: We also augment the test set with the spans that correspond to the concept. In particular, an example in the test set is of the form $(\mathbf{x}, e, \{\mathbf{s}_{i,e}\})$ where $\{\mathbf{s}_{i,e}\}$ is the set of spans that collectively identify the entity e in the text $\mathbf{x}$. In particular, each element in $\{\mathbf{s}_{i,e}\}$ encodes the character level beginning and end of the phrase in $\mathbf{x}$ that is constituent of e.

Thus, $\mathcal{D}_{train} = \{(\mathbf{x}_t, e)\}_{t=1}^{T}$ where $e \in \mathcal{E}_{seen}$ and $\mathcal{D}_{test} = \{(\mathbf{x}_k, e, \{\mathbf{s}_{i,e}\})\}_{k=1}^{K}$, where $e \in \mathcal{E}$.

		Seen		Unseen		Disjoint-Spans
		$\|\|\mathcal{E}_{\text{seen}}\|\|$	# Examples	$\|\|\mathcal{E}_{\text{unseen}}\|\|$	# Examples	Fraction of examples
RFE	Train	450	6430	n/a	n/a	unk
	Test	73	266	66	863	13%
hNLP	Train	1054	4377	n/a	n/a	5%
	Test	61	185	143	1018	7%

Table 2: Statistics of the datasets used in our experiments. Note that we do not need access to spans during training and therefore did not obtain span-level annotations for the RFE training set.

5.2. Dataset 1: Reason for Encounter

The Reason for Encounter (RFE) dataset is gathered from a telemedicine practice. Patients starting a visit describe their reason for seeking an encounter. This dataset contains a labeled subset of 4909 encounters with 4080 patients. The language used in RFE is more colloquial and less standardized, featuring many disjoint spans for medical entities. Each RFE is labeled by medical experts with corresponding medical findings using UMLS ontology. The RFE examples have an average length of 26 words.

We constructed the train-test dataset as outlined in § 5.1. In particular, $|\mathcal{E}_{seen}| = 450$ and $|\mathcal{E}_{unseen}| = 73$. This results in roughly 90% of the RFE dataset having at least one entity that is *seen*. Further, 24% of the examples in the RFE dataset have at least one entity in $\mathcal{E}_{unseen}$, and 10% of examples in the RFE dataset have all their entities in $\mathcal{E}_{unseen}$. We also provide the demographic breakdown of this dataset in Appendix A.1.

5.3. Dataset 2: hNLP dataset

Our second dataset is derived from the training data in the SemEval-2015 Task 14 (Elhadad et al., 2015). In particular, we start with the provided 136 discharge notes and their corresponding medical concepts along with their location spans. We split each discharge note into smaller text chunks using the newline delimiter. We removed chunks that do not have any entities associated with them. This leads to 5508 text chunks with an average length of 69.08 words. We built an initial dataset with text chunks, their entities, and their spans. These entities are UMLS Concept Unique Identifiers (CUIs).

We then constructed the train-test dataset as outlined in § 5.1. $|\mathcal{E}_{seen}| = 1054$ and $|\mathcal{E}_{unseen}| = 143$. This results in roughly 90% of the examples having at least one entity that is *seen*. For more detailed statistics on the dataset, see Table 2. For all examples in the test set, we attach the corresponding spans provided in the original dataset. We do not use these spans during training.

6. Task 1: Span Extraction

In this section, we describe the experiments for the task of entity span extraction using the OSLAT model described in §3.2.

6.1. Set-up

For entity span extraction, we compute the entity-attention scores for the ground-truth entities present in each input text. For experiments on both datasets, we compute the average entity-attention scores across all synonym terms associated with each ground-truth entity (identified by a UMLS CUI) as the exact matching synonym is not provided

Dataset	hNLP		RFE	
	Contiguous-Span s/u/all	Disjoint-Span s/u/all	Contiguous-Span s/u/all	Disjoint-Span s/u/all
Rule-Based	- / - / .74	- / - / .41	- / - / .55	- / - / .23
Fuzzy-Match	- / - / **.79**	- / - / .30	- / - / .35	- / - / .20
OSLAT	.77/**.73**/.74	- /**.47**/.47	**.67**/**.59**/**.66**	**.56**/**.60**/**.57**
OSLAT (CD)	**.80**/.65/.69	**.56**/.43/**.53**	.63/.52/.57	.52/.41/.45
OSLAT (NP)	.02/.02/.02	.00/.00/.00	.12/.11/.12	.05/.03/.05

Table 3: Micro-F1 scores for entity span extraction on both datasets, broken down by spans as well as examples with *seen* (s) and *unseen* (u) entities. We do not report separate seen and unseen values for the baselines since they are provided ground truth entities during inference. All reported results are averaged across 3 seeds, with $\sigma \leq 0.01$.

in the annotation. Since the attention scores are normalized through the softmax operation (sum up to 1), we manually set the threshold to be 0.05 during inference. Lastly, we also remove stop-words and punctuation marks from the predictions.

6.2. Metrics

We use the per-token micro-F1 score as the primary metric for evaluating our models for entity span extraction. This is done by computing the per-token precision and recall based on the token overlaps between the predicted and ground-truth spans before averaging across all examples.

6.3. Baselines

We compare **OSLAT** with the following baselines:

1. **Rule-based.** This is an in-house developed lookup-based approach that uses a sliding window strategy to find maximal matches of text corresponding to the entities and their synonyms. It ignores stop words while doing the match.
2. **Fuzzy-Match.** We adopt the fuzzy-string matching from the implementation by RapidFuzz (Bachmann, 2021), where spans with normalized Levenshtein Distance (Levenshtein, 1966) greater than a threshold are extracted for each entity.
3. **OSLAT (NP).** Ablation of OSLAT without self-alignment pretraining.
4. **OSLAT (CD).** Cross-dataset evaluation (model trained on RFE while evaluated on hNLP and vice versa).

Rule-based and fuzzy-match baselines are particularly strong because they are provided with the target entity and only need to string match one of the known entity synonyms to the target text. In particular, we find that these two baselines have very high precision, since the matched synonym is almost always the correct span.

6.4. Results

Table 3 shows the micro-F1 scores. We find that on the more challenging RFE dataset, OSLAT achieves the best performance. Even on the hNLP dataset, the RFE-trained OSLAT (OSLAT CD for hNLP) performs best on disjoint-spans. We believe that this is because the higher number of disjoint and otherwise complicated spans (where the

entity synonyms do not directly match the span text) in the RFE dataset force the model to learn more abstract label-attention representations. Note that this also demonstrates the generalizability of OSLAT, as there is a low entity overlap between the two datasets (Appendix A.2).

For both datasets, we also find that our model performs well for entities unseen during training. Since the synonym set often contains paraphrases of the same entity (*e.g.* stuffy nose, clogged nose), we hypothesize that our model learns to interpolate within the entity representation space and generalize to paraphrases for unseen entities.

7. Task 2 Results: Entity Linking

In this section, we describe the experiments for the task of entity linking using the approach described in §3.3.

7.1. Set-up

For entity linking, we cache the label representation for all entities $e \in \mathcal{E}$, before computing the retrieval score between all text-entity pairs. In practice, the retrieval score for each pair $(x_t, e_t) \in \mathcal{D}$ is computed as the max over all of e_t's synonyms $P(t)$, s.t. $s(x_t, e_t) = \arg\max_{p \in P(t)} \; s(x_t, p)$.

7.2. Metrics

To evaluate our entity linker, we use the top-k accuracy with $k = 1, 5, 10$. Specifically, for each ground-truth entity mentioned in the target text, we check if the text-entity pair has a top-k retrieval score.

7.3. Baselines

We use the bi-encoder retrieval architecture as a baseline, where we train BioBERT using the approach in Liu et al. (2021a). We first align the entity synonyms in the embedding space, before using the similarity between the target text and entity representation as the retrieval score.

We compare the following baselines:

1. **Fuzzy-Match.** Similar to § 6.3. We use Levenshtein Distance to score.
2. **BioBERT (Unsup)**. BioBERT bi-encoder directly as an unsupervised linker
3. **BioBERT (MS)**. BioBERT bi-encoder trained with Multi-Similarity (MS) loss. (Wang et al., 2019; Liu et al., 2021a)
4. **BioBERT (NCE)**. BioBERT bi-encoder trained with Noise Contrastive Estimation (InfoNCE) loss (Chen et al., 2020; Khosla et al., 2020; Gao et al., 2021).
5. **OSLAT (NP).** Ablation without self-alignment pertaining (no stage 1).
6. **OSLAT (No LA).** Ablation without label attention training (no stage 2).
7. **OSLAT (CD).** Cross-dataset evaluation (i.e. RFE model evaluated on hNLP and vice versa).

7.4. Results

Table 4 shows experiment results on entity linking. OSLAT outperforms all baselines on the RFE dataset, while lagging behind **BioBERT (NCE)** for @1 and @5 on the disjoint subset of hNLP. We hypothesize that the linking classifier might be struggling with the higher number of entities in the hNLP test set (See Table 2). As the number of entities increases, there exists a higher chance for semantically similar entities (*e.g.* back pain vs chronic back pain). Since the classifier only has access to the entity span representation $c^{(x_t, e)}$, which is a weighted sum of the

Dataset	hNLP		RFE	
Accuracy	Contiguous-Span @1/@5/@10	Disjoint-Span @1/@5/@10	Contiguous-Span @1/@5/@10	Disjoint-Span @1/@5/@10
Fuzzy-Match	**.289/.731/.840**	.018/.140/.228	.482/.740/.784	.123/.477/.554
BioBERT (Unsup)	.080/.134/.168	.070/.105/.211	.186/.031/.352	.077/.154/.185
BioBERT (MS)	.173/.280/.320	.105/.246/.333	.509/.734/.788	.339/.600/.723
BioBERT (NCE)	.198/.374/.455	**.123/.298**/.456	.467/.686/.776	.415/.692/.831
OSLAT	.224/.563/.713	.018/.193/**.491**	**.546/.865/.943**	**.554/.877/.954**
OSLAT (CD)	.238/.450/.577	**.123**/.193/.351	.510/.778/.858	.308/.646/.785
OSLAT (NP)	.001/.016/.028	.000/.000/.000	.004/.009/.019	.015/.015/.015
OSLAT (No LA)	.041/.071/.105	035/.053/.070	.070/.189/.271	.015/.138/.246

Table 4: Results for entity linking on both datasets, broken down by spans, and evaluated using top-k accuracy (@1, @5, @10).

encoder hidden states, it can be difficult to distinguish similar entities without explicitly mining for negatives. BioBERT bi-encoder retriever, on the other hand, has the full contextualized representation of the target text, and therefore may be able to better disambiguate similar entities if the discriminating context can be captured. We include the results for other encoder baselines in Appendix E.

8. Discussion

We propose OSLAT – a new architecture for entity span extraction and linking. OSLAT augments the standard label attention transformer architecture to allow an open set of labels. We also introduce a two-step training procedure including a modified supervised contrastive loss function, which we call the synonym-supervised NT-Xent loss.

In our experiments, we show that the two-step pretraining is critical, as both span extraction (§ 6) and entity linking (§ 7) task fail to be learned without the self-alignment pretraining of the encoder. This is because without pretraining, the encoder cannot meaningfully distinguish between the synonym and non-synonym representations of entities.

Through our detailed experiments, we first show that OSLAT implicitly learns to perform span extraction despite being trained only on text-level labels (without any span annotations). In entity linking, we also find that OSLAT outperforms other bi-encoder retrieval architectures. Most importantly, OSLAT performs well on entities mentioned in disjoint spans, which are very common in the colloquial language generated by patients.

Ethics This work was done as part of a quality improvement activity as defined in 45CFR §46.104(d)(4)(iii) – secondary research for which consent is not required for the purposes of "health care operations". In the "RFE" dataset, all ground truth annotations were performed by medical professionals who are full-time employees of the company.

References

Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17.

American Medical Informatics Association, 2001.

Max Bachmann. maxbachmann/RapidFuzz: Release 1.8.0. October 2021. doi: 10.5281/zenodo.5584996. URL https://doi.org/10.5281/zenodo.5584996.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. Semeval-2016 task 12: Clinical tempeval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California, June 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S16-1165.

Olivier Bodenreider. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32, 2004. ISSN 03051048. doi: 10.1093/nar/gkh061.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. Semantic retuning with contrastive tension. *International Conference on Learning Representations*, 7, 2021.

W W Chapman, W Bridewell, P Hanbury, G F Cooper, and B G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5): 301–10, Oct 2001.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Rhys Compton, Ilya Valmianski, Li Deng, Costa Huang, Namit Katariya, Xavier Amatriain, and Anitha Kannan. MEDCOD: A medically-accurate, emotive, diverse, and controllable dialog system. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 110–129. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/compton21a.html.

Koby Crammer, Mark Dredze, Kuzman Ganchev, Partha Pratim Talukdar, and Steven Carroll. Automatic code assignment to medical text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, BioNLP '07, page 129–136, USA, 2007. Association for Computational Linguistics.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. Semeval-2015 task 14: Analysis of clinical text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310, Denver, Colorado, June 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/S15-2051.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

Carol Friedman, Stephen B Johnson, Bruce Forman, and Justin Starren. Architectural requirements for a multipurpose natural language processor in the clinical environment. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 347. American Medical Informatics Association, 1995.

Jason A. Fries, Ethan Steinberg, Saelig Khattar, Scott L. Fleming, José D. Posada, Alison Callahan, and Nigam H. Shah. Trove: Ontology-driven weak supervision for medical entity classification. *CoRR*, abs/2008.01972, 2020. URL https://arxiv.org/abs/2008.01972.

Jinlan Fu, Xuanjing Huang, and Pengfei Liu. Spanner: Named entity re-/recognition as span prediction. *CoRR*, abs/2106.00641, 2021. URL https://arxiv.org/abs/2106.00641.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552. URL https://aclanthology.org/2021.emnlp-main.552.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1), oct 2021. ISSN 2691-1957. doi: 10.1145/3458754. URL https://doi.org/10.1145/3458754.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10, 1966.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.733. URL https://aclanthology.org/2020.emnlp-main.733.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 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, June 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.334.

URL https://aclanthology.org/2021.naacl-main.334.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459, Online and Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.109. URL https://aclanthology.org/2021.emnlp-main.109.

Yang Liu, Hua Cheng, Russell Klopfer, Matthew R. Gormley, and Thomas Schaaf. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953, Online and Punta Cana, Dominican Republic, November 2021c. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.481. URL https://aclanthology.org/2021.emnlp-main.481.

Veena Mayya, S. Sowmya Kamath, and Vijayan Sugumaran. $\mathcal{LATA}$ - label attention transformer architectures for icd-10 coding of unstructured clinical notes. In *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7, 2021. doi: 10.1109/CIBCB49929.2021.9562815.

Ali Mottaghi, Prathusha K. Sarma, Xavier Amatriain, Serena Yeung, and Anitha Kannan. Medical symptom recognition from patient text: An active learning approach for long-tailed multilabel distributions. *CoRR*, abs/2011.06874, 2020. URL https://arxiv.org/abs/2011.06874.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1100. URL https://aclanthology.org/N18-1100.

Bruce Nguyen and Shaoxiong Ji. Fine-tuning pretrained language models with label attention for biomedical text classification, 2021. URL https://arxiv.org/abs/2108.11809.

Naveen S. Pagad and N. Pradeep. Clinical named entity recognition methods: An overview. In Ashish Khanna, Deepak Gupta, Siddhartha Bhattacharyya, Aboul Ella Hassanien, Sameer Anand, and Ajay Jaiswal, editors, *International Conference on Innovative Computing and Communications*. Springer Singapore, 2022.

John P. Pestian, Chris Brew, Pawel Matykiewicz, DJ Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. A shared task involving multi-label classification of clinical free text. In *Biological, translational, and clinical language processing*, pages 97–104, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W07/W07-1013.

Viraj Prabhu, Anitha Kannan, Geoffrey J. Tso, Namit Katariya, Manish Chablani, David A. Sontag, and Xavier Amatriain. Open set medical diagnosis. *CoRR*,

abs/1910.02830, 2019. URL http://arxiv.org/abs/1910.02830.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

Özge Sevgili, Artem Shelmanov, Mikhail Y. Arkhipov, Alexander Panchenko, and Chris Biemann. Neural entity linking: A survey of models based on deep learning. *CoRR*, abs/2006.00575, 2020.

William F. Styler, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. de Groen, Bradley James Erickson, Timothy A. Miller, Chen Lin, Guergana K. Savova, and James Pustejovsky. Temporal annotation in the clinical domain. *TACL*, 2:143–154, 2014. URL https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/305.

Lynda Tamine and Lorraine Goeuriot. Semantic information retrieval on medical texts: Research challenges, survey, and open issues. *ACM Comput. Surv.*, 54 (7), sep 2021. ISSN 0360-0300. doi: 10.1145/3462476. URL https://doi.org/10.1145/3462476.

Ilya Valmianski, Nave Frost, Navdeep Sood, Yang Wang, Baodong Liu, James J. Zhu, Sunil Karumuri, Ian M. Finn, and Daniel S. Zisook. Smarttriage: A system for personalized patient data capture, documentation generation, and decision support. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 75–96. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/valmianski21a.html.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. A label attention model for icd coding from clinical text. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3335–3341, 2021.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Zero-shot entity linking with dense entity retrieval. *CoRR*, abs/1911.03814, 2019. URL http://arxiv.org/abs/1911.03814.

Appendix A. Dataset Statistics

A.1. RFE Demographics

The distribution of biological sexes in the dataset is 75% female and 25% male, the distribution of ages is 74% below 30 years old, 20% between 30 and 50 years old, and 6% above 50 years old. This distribution is not a random sample representative of the overall practice's population, but rather comes from a mixture of random samples drawn from two distinct times, and also from an active learning experiment for a different project.

A.2. Comparison

We found that there is a significant difference between the entity sets in both datasets (roughly 85% from hNLP to RFE and 69% from RFE to hNLP), although hNLP has twice the number of entities as the RFE dataset. We attribute the difference between the two datasets to their source; while RFE is derived from a telemedicine practice, hNLP is built from doctor's notes from in-patient settings. Second, only a tiny fraction of *unseen* entities in one dataset is *seen* in the other. This gives the assurance that when we evaluate the cross-domain task, we do not provide undue advantage to the model trained on the other dataset just because these *unseen* entities are known to the other dataset.

		RFE			**hNLP**		
		S	U	D	S	U	D
RFE	S	1	0	0	.23	.05	.72
	U	0	1	0	.09	.24	.67
hNLP	S	.10	.02	.88	1	0	0
	U	.12	.04	.84	0	1	0

Table A.1: Comparison of entities overlap between the two datasets.

In Table A.1, we quantitatively compare the overlap of entities between the datasets and make two observations. For each dataset (represented by rows), we present the number of entities in the *seen* training set (S), and in the *unseen* open set (U). In the columns corresponding to the other dataset, we provide the distribution of the occurrence of these entities in their *seen* (S) and *unseen* (U) concept distribution. The last column (D) corresponds to the proportion of concepts not represented in the other dataset (disjoint).

RFE
pregnancy, headache, dysuria, cough, abdominal pain, nausea, throat pain, UTI, delayed menstruation, vaginal pruritus, vaginal spotting, fever, crampy abdominal pain, fatigue, vomiting
hNLP
systemic arterial hypertension, edema, chest pain, coronary artery disease, pain, dyspnea, atrial fibrillation, heart failure, nausea, vomiting, bleeding, intracerebral hemorrhage, pneumonia, cyanosis, diabetes mellitus

Table A.2: Top 15 most frequent entities found in the two datasets.

This is also evident when we look at the top frequent entities from these two datasets in Table A.2 where hNLP focuses on more severe health issues (such as heart-related) that require hospitalization while RFE dataset focuses on non-urgent primary care services. However, they also share entities such as "vomiting."

We also found that only a tiny fraction of *unseen* entities in one dataset is *seen* in the other. This gives the assurance that when we evaluate the cross-domain task (§ 6.4) we do not provide undue advantage to the model trained on the other dataset just because these *unseen* entities are known to the

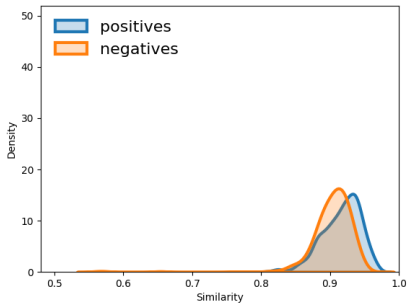
other dataset. Note that we did not intentionally construct the datasets this way and this result is a natural consequence of the significant difference in the vocabulary of the two datasets.

Appendix B. Training Hyperparameters

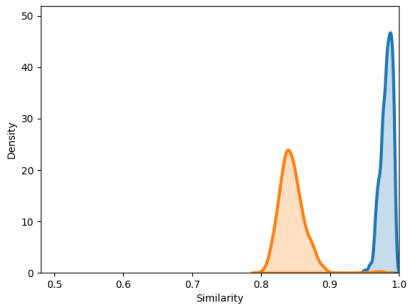
For both self-alignment pretraining (§ 3.1) and label attention training (§ 3.2), we use the ADAM optimizer (Kingma and Ba, 2014) with exponential decay after 1/10 of total steps and an effective batch size of 32. For self-alignment pretraining, we train the model for a total of 20 epochs with a learning rate of $2e-3$ and the number of negatives set to 50. For label attention training, we train for a total 10 epochs with a learning rate of $2e-4$ with the number of negatives set to 100. We set temperature τ to 0.07 based on the settings reported by Khosla et al. (2020). For training the classifier, we also train for a total of 10 epochs with a learning rate of $2e-4$ and the number of negatives set to 100. We follow the hyperparameters settings described in Lin et al. (2017), where $\alpha = 0.25$ and $\gamma = 2.0$

Appendix C. Effects of Self-Alignment Pretraining

To visualize the decrease in representational anisotropy, we plot the similarity between 1000 positive (synonyms) and negative (non-synonyms) entity pairs randomly sampled from $\mathcal{E}_{seen}$ (RFE). From Figure C.1, we see that before pretraining (a) the encoder could not differentiate representations of entity synonyms from non-synonyms, while after the pretraining (b), there is a dramatic shift that fully separates synonyms from non-synonyms.



(*a*) Before self-alignment pretraining



(*b*) After self-alignment pretraining

Figure C.1: Density plot of similarities between 1000 positive and negative entity pairs randomly sampled from $\mathcal{E}_{seen}$ (RFE).

Appendix D. Detailed metrics breakdown

In this section, we provide a detailed breakdown of the results from Table 3, where we discuss the recall-precision trade-off between our models and the two baseline methods. From the results in Table D.1, we see that while the RFE trained OSLAT achieved higher recall against both baseline methods, the rule-based model achieved higher precision across all datasets, with near-perfect precision for contiguous span entities. This is expected since the rule-based model has access to the ground-truth entity, the predictions it makes almost always exactly match with the entity or one of its synonyms. On the other hand, OSLAT can extract im-

Dataset	Entities	Metric (Micro)	OSLAT (RFE)	OSLAT (hNLP)	Rule-Based	Fuzzy
RFE Continuous-Span	seen	Precision	0.69±0.01	0.62±0.01	-	-
		Recall	0.65±0.00	0.69±0.01		
	unseen	Precision	0.59±0.01	0.53±0.01		
		Recall	0.59±0.01	0.50±0.01		
	all	Precision	0.67±0.01	0.57±0.01	**0.98**	0.90
		Recall	**0.64±0.00**	0.58±0.01	0.38	0.21
RFE Disjoint-Span	seen	Precision	0.61±0.01	0.60±0.01	-	-
		Recall	0.51±0.02	0.54±0.01		
	unseen	Precision	0.62±0.02	0.51±0.01		
		Recall	0.58±0.01	0.38±0.01		
	all	Precision	0.61±0.01	0.54±0.01	**0.95**	0.64
		Recall	**0.53±0.02**	0.44±0.01	0.12	0.12
hNLP Continuous-Span	seen	Precision	0.67±0.00	0.66±0.02	-	-
		Recall	0.97±0.00	0.92±0.01		
	unseen	Precision	0.52±0.01	0.61±0.01		
		Recall	0.88±0.01	0.90±0.00		
	all	Precision	0.57±0.01	0.61±0.01	**0.98**	0.70
		Recall	**0.91±0.01**	0.90±0.01	0.64	0.89
hNLP Disjoint-Span	seen	Precision	0.47±0.02	-	-	-
		Recall	0.71±0.02			
	unseen	Precision	0.43±0.02	0.45±0.01		
		Recall	0.45±0.02	0.47±0.01		
	all	Precision	0.44±0.02	0.45±0.01	**0.72**	0.49
		Recall	**0.51±0.02**	0.47±0.01	0.33	0.32

Table D.1: The breakdown of the micro-precision and recall performance on both datasets. We report the results for both of our models and the two baseline methods along with the standard deviation across 5 random seeds.

plicitly mentioned entities and disjoint-spans based on semantic similarity, resulting in a higher recall across all datasets. We leave the exploration of ensembling the two methods as a potential direction for future work. Lastly, it is worth mentioning that the precision and recall trade-off for OSLAT could be manually adjusted by tuning the prediction threshold of the attention scores. However, due to the limited size of our training set, we only report the performance for a fixed threshold (0.05).

Appendix E. Extended Results for Entity Linking

In Table E.1, we report the extended results for entity linking including the baselines of SAP-BERT (Liu et al., 2021a) and PubMedBERT (Gu et al., 2021). However, they are not directly comparable with **OSLAT**, since our model is based on BioBERT (Lee et al., 2020). We leave the experiments of OSLAT with other encoders as future work. Lastly, we also include the results using the ground-truth entity spans during inference (**OSLAT (GT)**). This is done by mean-pooling over the hidden states associated with the entity mention spans (rather than using the attention scores), before applying the binary classifier for prediction.

Dataset	hNLP		RFE	
	Contiguous-Span	Disjoint-Span	Contiguous-Span	Disjoint-Span
Accuracy	@1/@5/@10	@1/@5/@10	@1/@5/@10	@1/@5/@10
BioBERT (Unsup)	.080/.134/.168	.070/.105/.211	.186/.031/.352	.077/.154/.185
BioBERT (MS)	.173/.280/.320	.105/.246/.333	.509/.734/.788	.339/.600/.723
BioBERT (NCE)	.198/.374/.455	.123/.298/.456	.467/.686/.776	.415/.692/.831
SAP-BERT (Unsup)	.184/.297/.358	.193/.333/.404	.269/.403/.470	.108/.292/.354
SAP-BERT (MS)	.157/.234/.271	.088/.193/.211	.477/.705/.769	.354/.585/.615
SAP-BERT (NCE)	.206/.359/.435	.123/.351/.474	.197/.304/.351	.092/.185/.262
PubMedBERT (Unsup)	.080/.134/.168	.070/.105/.211	.144/.206/.234	.077/.139/.154
PubMedBERT (MS)	.197/.313/.363	.105/.246/.316	.523/.751/.820	.354/.692/.800
PubMedBERT (NCE)	.201/.379/.494	.175/.351/.561	.318/.447/.517	.200/.431/.539
OSLAT	.224/.563/.713	.018/.193/.491	.546/**.865**/**.943**	**.554**/**.877**/**.954**
OSLAT (CD)	.238/.450/.577	.123/.193/.351	.510/.778/.858	.308/.646/.785
OSLAT (NP)	.001/.016/.028	.000/.000/.000	.004/.009/.019	.015/.015/.015
OSLAT (No LA)	.041/.071/.105	035/.053/.070	.070/.189/.271	.015/.138/.246
OSLAT (GT)	**.483**/**.629**/**.752**	**.439**/**.597**/**.737**	**.555**/.793/.871	.462/.785/.908

Table E.1: Results for entity linking on both datasets, broken down by spans, and evaluated using top-k accuracy (@1, @5, @10).