MEDTYPE: Improving Medical Entity Linking with Semantic Type Prediction

Shikhar Vashishth[1] **Rishabh Joshi**[1] **Denis Newman-Griffis**[2]
Ritam Dutt[1] **Carolyn Rose**[1]
[1]Carnegie Mellon University [2]The Ohio State University
{svashish,rjoshi2,rdutt,cprose}@cs.cmu.edu
newman-griffis.1@osu.edu

arXiv:2005.00460v2 [cs.CL] 16 Sep 2020

Abstract

Medical entity linking is the task of identifying and standardizing medical concepts referred to in an unstructured text. Most of the existing methods adopt a three-step approach of (1) detecting mentions, (2) generating a list of candidate concepts, and finally (3) picking the best concept among them. In this paper, we probe into alleviating the problem of overgeneration of candidate concepts in the candidate generation module, the most under-studied component of medical entity linking. For this, we present MEDTYPE, a fully modular system that prunes out irrelevant candidate concepts based on the predicted *semantic type* of an entity mention. We incorporate MEDTYPE into five off-the-shelf toolkits for medical entity linking and demonstrate that it consistently improves entity linking performance across several benchmark datasets. To address the dearth of annotated training data for medical entity linking, we present WIKIMED and PUBMEDDS, two large-scale medical entity linking datasets, and demonstrate that pre-training MEDTYPE on these datasets further improves entity linking performance. We make our source code and datasets publicly available for medical entity linking research.

1 Introduction

Identifying the standardized concepts referred to in an unstructured text is a critical component of biomedical natural language processing, enabling harmonization across different documents for search and semantic analysis (Jovanović and Bagheri, 2017). Medical entity linking, also referred to as medical concept normalization, is the task of harmonizing different surface forms for the same concepts, for example, so that documents mentioning *Amyotrophic lateral sclerosis* and *Lou Gehrig's Disease* can be correctly analyzed as referring to the same disease. Medical entity linking involves three distinct subproblems: (1) *named entity recognition* (NER), identifying the mentions of medical concepts of interest; (2) *candidate generation*, identifying a set of standardized concepts that a specific mention may refer to; and (3) *word/phrase sense disambiguation* (WSD), picking the best candidate concept for the observed context.

Compared to NER and WSD, candidate generation is an under-studied component of medical entity linking. Prior methods have historically relied on dictionary lookup and string matching (Aronson and Lang, 2010; Savova et al., 2010) for both NER and candidate generation, an approach yielding high precision but incomplete coverage (Travers and Haas, 2006; Reátegui and Ratté, 2018). Recent neural methods have taken the opposite approach and included all concepts in a fixed subset as candidates (Tutubalina et al., 2018; Soysal et al., 2018; Zhao et al., 2019; Mondal et al., 2019), which rapidly becomes intractable when generalizing to wider-coverage concept vocabularies. For example, the full Unified Medical Language System (UMLS) (Bodenreider, 2004), a large-scale resource combining over 150 biomedical terminologies, contains over 4 million unique concepts. Thus, robust strategies to reduce *overgeneration* of candidates are key to leveraging the high coverage afforded by neural approaches.

In this paper, we present MEDTYPE, a fully modular system for pruning out overgenerated candidates in medical entity linking by predicting the *semantic type* of an entity mention. We make the following contributions:

- We incorporate MEDTYPE into five off-the-shelf toolkits for medical entity linking and demonstrate that it consistently improves entity linking performance across several benchmark datasets.
- To address the dearth of annotated training data

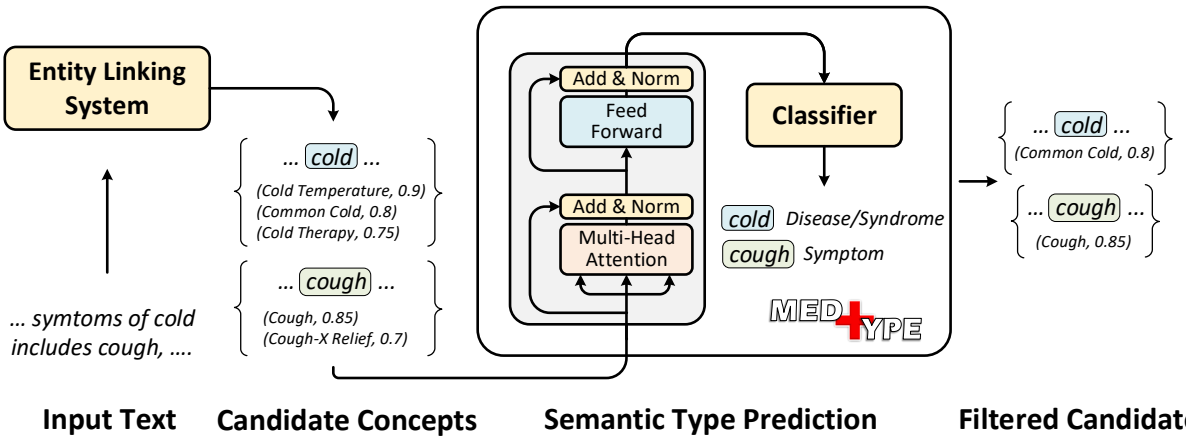


Figure 1: Overview of MEDTYPE. For a given input text, MEDTYPE takes in the set of identified mentions along with their list of candidate concepts as input. Then, for each mention MEDTYPE predicts its semantic type based on its context in the text. The identified semantic type is used to filter out the irrelevant candidate concepts thus controlling overgeneration of candidates and improving medical entity linking. Please refer to Section 4 for details.

for medical entity linking, we present WIKIMED and PUBMEDDS, two large-scale medical entity linking datasets, and demonstrate that pretraining MEDTYPE on these datasets further improves entity linking performance.

- We show that type-based filtering significantly reduces the number of candidates for WSD, enabling further improvements in the final step of medical entity linking.

MEDTYPE's source code and datasets proposed in the paper are included as supplementary material.

2 Related Work

Identifying named entity mentions and linking them to concepts in a knowledge base is an essential task for semantic understanding of natural text and information extraction. In addition to jointly modeling NER and entity linking as interdependent tasks (Kolitsas et al., 2018; Martins et al., 2019), many studies leverage the rich semantics of the target knowledge base to improve linking (Radhakrishnan et al., 2018; Wu et al., 2020). Many knowledge bases group entities into semantic types, which is highly informative for many downstream NLP tasks such as co-reference resolution (Durrett and Klein, 2014), relation extraction (Yaghoobzadeh et al., 2017; Vashishth et al., 2018), question answering (Das et al., 2017), and language modeling (Zhang et al., 2019). Recent studies have shown that fine-grained entity type prediction improves entity linking over Wikipedia (Ling et al., 2015; Raiman and Raiman, 2018), indicating clear potential for type prediction as a standard component of entity linking pipelines.

Medical Entity Linking: In the medical domain, the role of entity typing in selecting suitable candidates for medical concept mentions was recognized in some of the earliest rule-based medical entity extraction tools (Aronson et al., 1994). However, type prediction is typically deeply embedded in rule-based tools, reducing generalizability, and its use in deep learning systems has been minimal. Nejadgholi et al. (2019) utilized neural language modeling frameworks to identify the semantic type of a mention in a medical text, but do not apply their predictions downstream; by contrast, Loureiro and Jorge (2020) utilized approximate dictionary matching heuristics with specialized neural language models to improve both medical entity typing and entity linking in biomedical literature. However, these works have not explored the efficacy of utilizing type information within the entity linking task itself. This work alleviates the overgeneration problem produced by both rule-based (Reátegui and Ratté, 2018) and deep learning systems in practice, by using predicted semantic type to prune irrelevant candidates, and does so in a modular approach making it easy to incorporate in any entity linking architecture.

3 MEDTYPE Overview

In the medical domain, the inability to resolve ambiguity among possible concepts for a given mention has been identified as a pronounced weakness (Aronson and Lang, 2010) of existing entity linking systems. For instance, a mention *'cold'* can refer to distinct concepts such as *common cold* (disease),

cold temperature (natural phenomena), or *cold brand of chlorpheniramine-phenylpropanolamine* (pharmacologic substance). This ambiguity because of *polysemy* and *homonymy* leads to overgeneration of candidate concepts. Thus, including an additional step for pruning irrelevant candidates has the potential to improve entity linking performance.

In this work, we propose MEDTYPE, which takes in a set of generated candidates for a mention and filters them based on the predicted semantic type. Most of the existing medical entity linkers (Savova et al., 2010; Aronson and Lang, 2010; Neumann et al., 2019) are non-neural and predominantly rely on substring matching for linking mentions to entities. However, MEDTYPE utilizes recent advances in deep learning-based language modeling techniques (Peters et al., 2018; Devlin et al., 2018) for encoding context to predict the semantic type of a mention. The overall architecture of MEDTYPE is shown in Figure 1. Further details of the method are presented in Section 4. Apart from that, we also propose two large-scale annotated datasets: WIKIMED and PUBMEDDS, which can serve as a great resource for the community. The proposed datasets are around 3× and 164× times larger than the current largest medical entity linking dataset MedMentions (Mohan and Li, 2019). More details on datasets creation are provided in Section 5.

4 MEDTYPE Details

Formally, the entity linking task is defined as: Let $\mathcal{E} = \{e_1, e_2, ..., e_N\}$ be a predefined set of entities in a knowledge graph and $\mathcal{T} = (w_1, w_2, ..., w_{|\mathcal{T}|})$ be a given unstructured text with n tokens. The entity linking task involves identifying mentions m of the form $w_{i...j}$ in $\mathcal{T}$ and mapping them to an entity $e \in \mathcal{E}$. Following prior works (Soldaini and Goharian, 2016; Neumann et al., 2019), we define $\mathcal{E}$ as the set of entities in the Unified Medical Language System, or UMLS (Bodenreider, 2004), a large-scale compendium of multiple source vocabularies which provides broad coverage of over 4 million biomedical concepts. Most of the existing entity linking methods follow a three-step procedure: (1) Named Entity Recognition (NER), the task of recognizing concept mentions $\{m_1, m_2, ..., m_k\}$ in the text, (2) Candidate Generation, involves generating a probable set of candidates $\mathcal{C}_i = \{e_1^i, e_2^i, ..., e_l^i \mid e_j^i \in \mathcal{E}\}$ for each identified mention m_i, and finally (3) Word/phrase sense disambiguation (WSD) for assigning a score $s_j^i \in \mathbb{R}$ for each candidate concept $e_j^i \in \mathcal{C}_i$.

As most of the existing techniques are based on dictionary lookup and sub-string matching (Aronson and Lang, 2010; Savova et al., 2010), the sheer number of medical concepts and prominence of lexical ambiguity among mentions due to *homonymy* and *polysemy* (Weeber et al., 2001; Savova et al., 2008) leads to overgeneration of candidate concepts. For alleviating this problem, we propose MEDTYPE, a fully modular system that takes in a generated candidate set $\mathcal{C}$ for a given mention and outputs a filtered set $\mathcal{C}' \subseteq \mathcal{C}$ as depicted in Figure 1. This is attained by predicting the *semantic type* of the mention based on its context. UMLS provides a mapping of its concepts to one or more semantic types such as *Disease or Syndrome*, *Anatomical Structure*, etc., out of a total of 127 types. MEDTYPE thus models entity typing as a multi-label classification problem, and utilizes the predicted types of a given mention for pruning the candidate concept set. As many of these types are very fine-grained and have sparse coverage in entity linking corpora, we use the coarse-grained Semantic Groups developed by (McCray et al., 2001) and *is-a* relationships in the Semantic Network to re-label these types into 24 semantic groups.

Context Encoder: For training MEDTYPE, we take the input data of the form $\mathcal{D} = [(x_0, y_0), ..., (x_N, y_N)]$ where x_i denotes mention m_i along with its context which comprises of its neighboring tokens in a window of size k, i.e., $Con(m_i, k) = (m_i^{-k}, ..., m_i^{-1}, m_i^1, ..., m_i^k)$ and y_i is the semantic type. Motivated by the ability to handle polysemous tokens and superior modeling capabilities of long range dependencies of Transformer-based models (Vaswani et al., 2017), we utilize a pre-trained BERT (Devlin et al., 2018) encoder and fine-tune it for our type prediction task. We feed the mention concatenated with its context, i.e., $[m; Con(m_i, k)]$ as input to the encoder along with their positional information for allowing the model to distinguish between them. Finally, the embedding corresponding to the initial `[CLS]` token is passed to a feed-forward classifier for 24-way prediction of semantic types.

5 Training corpora

The availability of large scale public datasets has helped to foster research in various domains (Rus-

sakovsky et al., 2015; Bowman et al., 2015). However, obtaining and sharing medical data presents a significant obstacle, given the sensitive nature of the data and privacy issues (Cios and Moore, 2002). Moreover, crowd-sourced annotation of medical information requires expensive and difficult-to-obtain expertise to ensure data quality (Fries et al., 2019). We overcome these challenges by automatically creating two large training datasets for medical entity linking.

5.1 Transfer Learning with WIKIMED

In order to leverage transfer learning, an effective tool to solve the problem of insufficient training data in specialized domains (Wang and Mahadevan, 2011; Li et al., 2014; Zhou et al., 2014), we developed a Wikipedia-based dataset for medical entity linking, which we use to pre-train MEDTYPE before fine-tuning it on medical domain text. Wikipedia, though not a medical data source, does include many mentions of medical concepts that can inform entity typing models. We utilize three knowledge sources to map hyperlinks in Wikipedia articles to UMLS entities: Wikidata (Vrandečić and Krötzsch, 2014), Freebase (Bollacker et al., 2008), and the NCBI Taxonomy (Page, 2011). This maps approximately 60,500 Wikipedia page IDs to UMLS concepts, with which we extract around 1 million relevant mentions from 6 million Wikipedia articles to create the WIKIMED dataset. More data statistics of WIKIMED are summarized in Table 1.

Since Wikipedia is human-curated, we assume that most of its linkages are of sufficiently high-quality to yield accurate mappings. Thus, we also utilize WIKIMED as one of the evaluation datasets for assessing the performance of medical entity linkers on non-medical domain text. WIKIMED is an order of magnitude larger in size than previous medical entity linking datasets such as MedMentions (Mohan and Li, 2019) and the NCBI disease corpus (Doğan et al., 2014), and covers medical entities from diverse semantic types. We provide a detailed comparison of semantic types coverage of datasets in Appendix Table 5.

5.2 Distant Supervision in Biomedical Language: PUBMEDDS

Distant supervision (Mintz et al., 2009) enables automatic generation of training data and has been exploited for several tasks (Reschke et al., 2014; Fan et al., 2015; Purver and Battersby, 2012), including identifying potential mentions of medical concepts (Newman-Griffis et al., 2018). To create a large-scale training dataset for medical entity linking drawn from biomedical language, we use distant supervision on PubMed abstracts to generate PUBMEDDS. Unlike Wikipedia, PubMed abstracts do not include *a priori* links to database entities: we, therefore, run a state-of-the-art biomedical NER model (Neumann et al., 2019) on 20 million PubMed abstracts to extract medical entity mentions. We then make use of the Medical Subject Headings (MeSH) tags assigned to each PubMed article to filter the extracted entity mentions: a mention is retained only when it exactly matches with the name of one of the provided MeSH headers. The UMLS maps MeSH headers to UMLS concept identifiers, which we utilize to get the semantic type of each extracted mention. The procedure gives us PUBMEDDS, a dataset with 58 million annotated mentions which we utilize for pre-training MEDTYPE. In Section 7.2, we demonstrate that although PUBMEDDS is distantly-supervised, it has sufficiently high precision and can serve as an eminent resource for improving medical entity linking

Datasets	#Docs	#Sents	#Ment	#Unq Con
NCBI	792	7,645	6,817	1,638
Bio CDR	1,500	14,166	28,559	9,149
ShARe	431	27,246	17,809	1,719
MedMentions	4,392	42,602	352,496	34,724
WIKIMED	393,618	11,331,321	1,067,083	57,739
PUBMEDDS	13,197,430	127,670,590	57,943,354	44,881

Table 1: Details of the medical entity linking datasets used in our experiments. WIKIMED and PUBMEDDS are our novel automatically-created datasets.

6 Experimental Setup

6.1 Datasets

In our experiments, we evaluate the models on four benchmark datasets: NCBI (Doğan et al., 2014), Bio CDR (Li et al., 2016), ShARe (Elhadad et al., 2015), and MedMentions (Mohan and Li, 2019) for medical entity linking, as well as our novel WIKIMED dataset. Please note that we avoid evaluation on distantly-supervised PUBMEDDS as some of its mentions might be wrongly annotated (refer to Section 7.2). Overall, the chosen datasets span across different domains such as biomedical research articles, Electronic Health Records (EHR), and general domain text. Thus, allow us to evaluate the generality of MEDTYPE across diverse domains. The dataset statistics are presented in

	NCBI	Bio CDR	ShARe	MedMentions	WikiMed
MetaMap	7.7	19.9	7.9	36.7	64.3
Oracle (Fine)	+0.5	+0.2	+0.3	+6.4	+0.8
Oracle (Coarse)	+0.5	+0.2	+0.3	+5.7	+0.6
MedType	+0.4	+0.1	+0.2	+3.9	+0.5
cTAKES	18.0	40.0	18.2	16.9	10.4
Oracle (Fine)	+0.2	+0.2	+0.1	+0.2	+0.1
Oracle (Coarse)	+0.2	+0.2	+0.1	+0.2	+0.1
MedType	+0.2	+0.2	+0.1	+0.1	0.0
MetaMapLite	6.4	18.4	7.4	32.6	62.0
Oracle (Fine)	+1.6	+1.5	+1.5	+7.2	+1.9
Oracle (Coarse)	+1.6	+1.5	+1.5	+6.0	+1.6
MedType	+1.6	+1.4	+1.4	+4.2	+1.5
QuickUMLS	5.3	11.6	5.2	9.0	4.7
Oracle (Fine)	+0.7	+3.4	+2.7	+2.4	+1.9
Oracle (Coarse)	+0.7	+3.0	+2.7	+2.1	+1.6
MedType	+0.6	+2.9	+2.4	+1.3	+1.6
ScispaCy	11.5	22.6	9.1	37.2	41.1
Oracle (Fine)	+1.4	+1.5	+1.4	+8.2	+1.9
Oracle (Coarse)	+1.4	+1.5	+1.4	+6.8	+1.7
MedType	+1.4	+1.3	+1.3	+3.9	+1.5

Table 2: For quantifying the impact of MedType on medical entity linking, we report the F1-score for five medical entity linking methods on multiple datasets. For each method, the first row is its base performance, and the following rows indicate the change in F1-score on incorporating a type-based candidate concepts filtering step. We report the results with the oracle type predictors (fine-grained and coarse-grained) and MedType. Overall, we find that MedType gives performance comparable to an oracle and improves medical entity linking across all settings. Please refer to Section 7.1 for details.

Table 1. We provide a short description of each dataset in Appendix A.

6.2 Medical Entity Linkers

For evaluating the effectiveness of MedType, we utilize it for pruning candidate concepts generated from the following five entity linking models: MetaMap (Aronson and Lang, 2010), cTAKES (Savova et al., 2010), MetaMapLite (Demner-Fushman et al., 2017), QuickUMLS (Soldaini and Goharian, 2016), and ScispaCy (Neumann et al., 2019). To the best of our knowledge, the list covers almost all the existing entity linkers in the medical domain. A brief description of each method is included in Appendix B.

6.3 Evaluation Metrics

For evaluating performance on entity linking task, we report F1-score for the following three metrics: (1) *Exact_mention_id_match* where score is given only when there is exact match of the mention as well as the entity concept identifier. This is directly adopted from TAC KBP 2013[1]. (2) *Partial_mention_id_match* where we give partial score for mention and entity id match as proposed by Tong et al. (2018). (3) *Only_entity_id_match* where we consider only entity identifiers irrespective of the mentions for which it is identified. For computing all the scores, we use an open-source entity linking evaluation toolkit[2].

7 Results

In this section, we answer the following questions.

Q1. Does incorporating MedType in existing entity linking systems help? (Section 7.1)

Q2. How effective are proposed datasets for semantic type prediction? (Section 7.2)

Q3. Where does gains on using WikiMed and PubMedDS come from? (Section 7.3)

Q4. How much does type-based filtering help prune irrelevant candidates ? (Section 7.4)

7.1 Entity Linking with MedType

For demonstrating the effectiveness of MedType for medical entity linking, we compare its perfor-

[1]https://tac.nist.gov/2013/KBP/

[2]https://github.com/wikilinks/neleval

mance with entity linkers listed in Section 6.2. We report the results for *Exact_mention_id_match* as described in Section 6.3 and results with other metrics are included in Appendix Table 6 and 7. For WIKIMED dataset, we evaluate only on identified mentions which overlap with the ground truth. We make this exception since Wikipedia editors are encouraged to link only the early occurrences of an entity, as opposed to exhaustive annotations. Moreover, in WIKIMED, we only include the mentions for which mapping to UMLS is available, which is not comprehensive. For highlighting the maximum possible improvement from type-based candidate concepts pruning, we report the results with an *Oracle* model. As described in Section 4, in our experiments, we work with 24 coarse-grained semantic types instead of 127 types provided by UMLS. Hence, for comparison, we present the results with both fine-grained (*Oracle (Fine)*) and coarse-grained type (*Oracle (Coarse)*) oracles.

The overall results are presented in Table 2. For each entity linking system, we report its default performance and the change in scores when used along with different candidate concepts filtering methods. The results for MEDTYPE are obtained after training it on WIKIMED, PUBMEDDS, and the training split of the corresponding datasets. Overall, we find that across all the models and datasets, MEDTYPE gives a substantial improvement in performance. Moreover, in none of the settings, it leads to any degradation. Thus, including a type-based filtering step only enhances entity linking systems. Further, we note that the gain with MEDTYPE is comparable to improvement with using an oracle. The small improvement in the case of cTAKES model is attributed to the fact that it generates substantially fewer candidates for detected mentions. Thus, it does not provide much scope for pruning irrelevant candidates. We used paired bootstrap significance testing (Berg-Kirkpatrick et al., 2012) for validating statistical significance ($p < 0.01$) of improvements from MEDTYPE. The results overall justify the central thesis of this work, that pruning irrelevant candidate concepts based on semantic type helps improve medical entity linking.

7.2 Benefit of utilizing WIKIMED and PUBMEDDS

In this section, we analyze the impact of using our proposed datasets: WIKIMED and PUBMEDDS for semantic type prediction. Since we model type prediction as a multi-label classification problem, we use the area under the Precision-Recall curve (AUC) as our evaluation metric. For quantifying the benefit of utilizing our proposed datasets, we report the performance of MEDTYPE trained under different settings, as defined below.

	NCBI	Bio CDR	ShARe	MedMentions
Training data (T)	89.2	75.8	73.4	74.8
T ← WIKIMED	95.0	91.0	90.0	76.2
T ← PUBMEDDS	95.5	**95.7**	91.9	80.9
T ← Both	**96.5**	95.6	**93.1**	**81.7**

Table 3: Evaluating the effect of using WIKIMED and PUBMEDDS for semantic type prediction. We use the area under PR-curve as the evaluation metric. T ← X denotes MEDTYPE first trained on X dataset then fine-tuning using T. We find that datasets WIKIMED and PUBMEDDS provide a substantial improvement on the task. More details are provided in Section 7.2.

- **Training data (T)** denotes MEDTYPE trained on the training split of the corresponding datasets.
- **T ← WIKIMED** refers to the model first trained on WIKIMED and then fine-tuned using the training data.
- **T ← PUBMEDDS** similar to T ← WIKIMED, indicates MEDTYPE first trained on PUBMEDDS and then fine-tuned on the training data.
- **T ← Both** denotes the combined model which utilizes both the proposed datasets. It concatenates BERT encoding from T ← WIKIMED and T ← PUBMEDDS models and passes it to a classifier which is trained using the training dataset.

The overall results are presented in Table 3. We find that there is a substantial gain in performance on using WIKIMED and PUBMEDDS along with the training data. Further, we find that the combined model which allows to incorporate the benefits from both the corpora gives the best performance on 3 out of 4 datasets. This shows that both the datasets contain complementary information relevant for semantic type prediction. Overall, we get an average absolute increase of 9.8, 12.7, and 13.4 AUC from WIKIMED, PUBMEDDS, and the combined corpora respectively.

On Quality of Proposed Datasets: Automatic dataset creation is prone to a variety of errors. In order to assess the quality of the distantly supervised PUBMEDDS dataset, we identified the subset of documents overlapping with three manually-

	Precision	Recall
NCBI	86.3	6.5
Bio CDR	75.8	1.3
MedMentions	90.3	5.3

Table 4: Assessing the quality of proposed PUBMEDDS dataset by evaluating its precision and recall on biomedical documents common with NCBI, Bio CDR, and MedMentions. We find that although PUBMEDDS has low coverage, extracted mentions have on average 84% precision across the three datasets..

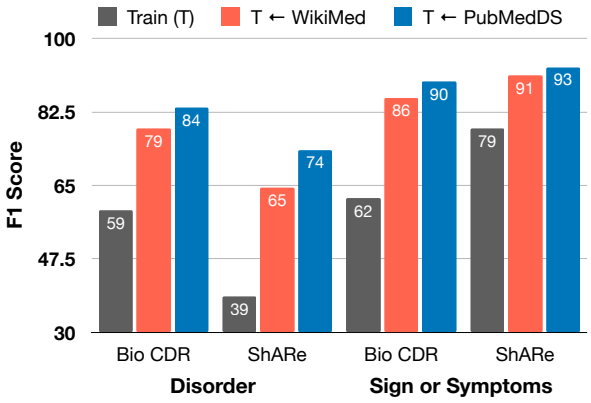


Figure 2: Analysis of the impact on using WIKIMED and PUBMEDDS datasets on two semantic types: *Disorder* and *Sign or Symptoms* on Bio CDR and ShARe datasets. We find that training MEDTYPE on our proposed datasets gives considerable improvement on semantic type prediction. Refer to Section 7.3 for details.

annotated datasets using PubMed abstracts: MedMentions, NCBI, and Bio CDR. We note that all the documents in the three datasets are part of PUBMEDDS. This allowed us to evaluate the precision and recall of our distantly supervised mentions with respect to human annotations. The results of this analysis are reported in Table 4. Reflecting on the strict requirements for keeping a mention in our dataset (identification with a NER tool and exact match to a provided MeSH header), we find that the coverage of all possible mentions in the document is low, but precision is relatively high at around 84%. Thus, the annotations that are provided in PUBMEDDS are of high quality for training entity linking models.

WIKIMED was created using human-provided links and expert-curated mappings, and can, therefore, be considered inherently high precision. However, Wikipedia editors are encouraged to only link early occurrences of an entity in the articles, as opposed to exhaustive annotations; thus, as with our findings in PUBMEDDS, we can consider WIKIMED to be relatively low in coverage of all true mentions in the documents. Nevertheless, our proposed datasets do provide a reliable signal for entity linking.

7.3 Type-wise Analyzing of Improvement from WIKIMED and PUBMEDDS

In this section, we investigate the cause behind the improvement on semantic type prediction from WIKIMED and PUBMEDDS datasets as discussed in the previous section. In Figure 2, we report the F1-score for two datasets: Bio CDR and ShARe, on which we obtain the highest gains in Table 3. The results are reported for T, T ← WIKIMED, and T ← PUBMEDDS models on prominent semantic types. We present the results across all semantic types on all the datasets in Appendix Table 8, 9, 10, and 11. The results show that for semantic types such as *Sign or Symptoms*, utilizing WIKIMED and PUBMEDDS give an absolute increase of 24, 28 on Bio CDR and 12, 14 on ShARe respectively. A more substantial increase is observed on *Disorder* semantic type which has relatively less coverage in the training split of the datasets. Thus, the results give insight behind the overall gain which we obtain on utilizing additional corpora.

7.4 Impact of Semantic Type Prediction on Candidate Generation

Table 2 demonstrates that MEDTYPE consistently improves entity linking performance when using the candidate scoring methods provided in each of our evaluated entity linkers. However, candidate ranking and disambiguation are active areas of research (Mondal et al., 2019; Pesaranghader et al., 2019), and MEDTYPE's modular nature makes it easy to incorporate into any entity extraction pipeline. We therefore investigated the impact of semantic type prediction in filtering out overgenerated candidate concepts in order to simplify the sense disambiguation task.

Semantic type-based pruning consistently reduces the candidate set size. Figure 3 illustrates the outcomes of type-based pruning on the candidate set sizes for the 38,349 samples in the MedMention test set where ScispaCy included the correct CUI in its candidate set. Oracle type information, representing the upper bound of what type-based pruning can achieve, reduces the candidate set size in over 80% of cases at the fine level, and directly solves the sense disambiguation problem in

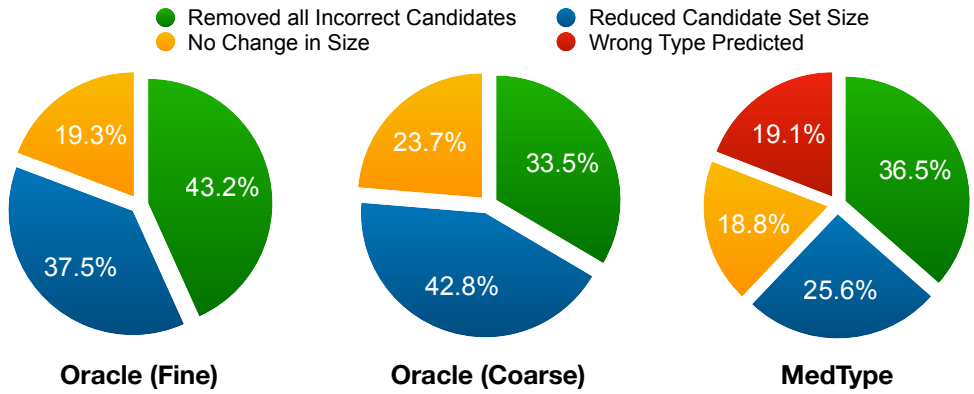


Figure 3: Outcomes of using MEDTYPE to prune candidate concept sets in MedMentions, using ScispaCy. Mentions for which ScispaCy did not generate one or more correct candidate are omitted. The results are thus for 38,234 annotated mentions in the dataset.

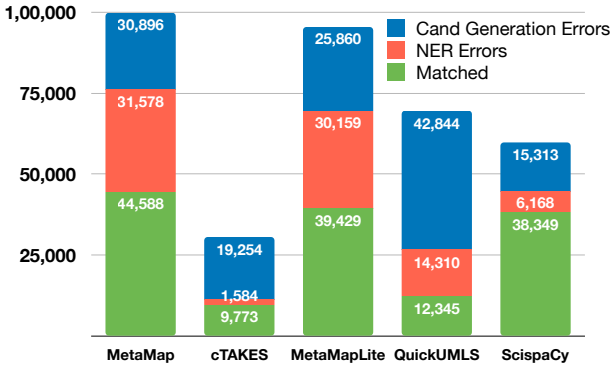


Figure 4: Error analysis of output predictions from all entity linkers on MedMentions test set (annotated set size: 70,405 mentions). NER Errors indicate false positive entity spans; Candidate Generation Errors indicate exclusion of the correct entity from the returned candidate list; Matched indicates that neither of these errors were present.

43% of cases. Coarse-grained typing only slightly reduces these gains while significantly simplifying the type prediction problem, further supporting our choice of coarse labels for MEDTYPE. MEDTYPE, in turn, achieves most of the reductions in candidate set size yielded by oracle information, and the performance improvements shown in Table 2 clearly demonstrate the practical gains from this filtering. Appendix Figure 5 provides further details on the degree of reduction in candidate set size.

MEDTYPE can help improve the full extraction pipeline. Failures can occur at all three stages of entity extraction: NER, candidate generation, and sense disambiguation. Figure 4 illustrates the number of medical entities recognized by the existing entity linkers in the MedMentions test set, broken down into (1) NER errors, where the linker produced a false positive entity span; (2) candidate generation errors, where the linker did not include the correct entity in the candidate list; and (3) matches, where the linker found a valid span and produced the correct entity. We see that the five linkers evaluated vary widely in the number of entities output, but in all cases include a significant number of both NER and candidate generation errors. In addition to MEDTYPE's utility in reducing candidate set sizes, which allows for broader-coverage candidate generation methods, we also observe that in all cases where a NER error was produced, MEDTYPE classified it as a *None* type; this indicates clear utility in incorporating MEDTYPE as a component of any system to filter out false positives in NER.

8 Conclusion

We presented MEDTYPE, a fully modular system for improving medical entity linking by filtering out irrelevant candidate concepts based on the predicted semantic type of an entity mention. MEDTYPE improves entity linking performance for a variety of popular medical entity extraction toolkits across several benchmark datasets, clearly demonstrating the utility of a type-based filtering step in medical entity linking. We further present two novel large-scale, automatically-created datasets of medical entity mentions: WIKIMED, a Wikipedia-based dataset for cross-domain transfer learning, and PUBMEDDS, a distantly-supervised dataset of medical entity mentions in biomedical abstracts. Pre-training on these datasets substantively improves MEDTYPE performance, and we share these datasets with the community as a resource for medical entity linking research.

References

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–36.

Alan R Aronson, Thomas C Rindflesch, and Allen C Browne. 1994. Exploiting a Large Thesaurus for Information Retrieval. In *Intelligent Multimedia Information Retrieval Systems and Management - Volume 1*, RIAO '94, pages 197–216, Paris, FRA. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):D267–D270.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA. ACM.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Krzysztof J. Cios and G. William Moore. 2002. Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, 26(1):1 – 24. Medical Data Mining and Knowledge Discovery.

Rajarshi Das, Manzil Zaheer, Siva Reddy, and Andrew McCallum. 2017. Question answering on knowledge bases and text using universal schema and memory networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 358–365, Vancouver, Canada. Association for Computational Linguistics.

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

Dina Demner-Fushman, Willie J Rogers, and Alan R Aronson. 2017. Metamap lite: an evaluation of a new java implementation of metamap. *Journal of the American Medical Informatics Association*, 24(4):841–844.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the association for computational linguistics*, 2:477–490.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. 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. Association for Computational Linguistics.

Miao Fan, Qiang Zhou, and Thomas Fang Zheng. 2015. Distant supervision for entity linking. *arXiv preprint arXiv:1505.03823*.

Jason A. Fries, Paroma Varma, Vincent S. Chen, Ke Xiao, Heliodoro Tejeda, Priyanka Saha, Jared Dunnmon, Henry Chubb, Shiraz Maskatia, Madalina Fiterau, Scott Delp, Euan Ashley, Christopher Ré, and James R. Priest. 2019. Weakly supervised classification of aortic valve malformations using unlabeled cardiac mri sequences. *Nature Communications*, 10(1):3111.

V. N. Garla and C. Brandt. 2012. Knowledge-based biomedical word sense disambiguation: an evaluation and application to clinical document classification. *Journal of the American Medical Informatics Association*, pages 882–886.

Vijay Garla, Vincent Lo Re, Zachariah Dorey-Stein, Farah Kidwai, Matthew Scotch, Julie Womack, Amy Justice, and Cynthia Brandt. 2011. The Yale cTAKES extensions for document classification: Architecture and application. *Journal of the American Medical Informatics Association*, 18(5):614–620.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Jelena Jovanović and Ebrahim Bagheri. 2017. Semantic annotation in biomedicine: The current landscape. *Journal of Biomedical Semantics*, 8(1):1–18.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, page arXiv:1412.6980.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

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

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 : the journal of biological databases and curation*, 2016:baw068.

W. Li, L. Duan, D. Xu, and I. W. Tsang. 2014. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1134–1148.

Xiao Ling, Sameer Singh, and Daniel S. Weld. 2015. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328.

Daniel Loureiro and Alípio Mário Jorge. 2020. Medlinker: Medical entity linking with neural representations and dictionary matching. In *Advances in Information Retrieval*, pages 230–237, Cham. Springer International Publishing.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. 2019. Joint learning of named entity recognition and entity linking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 190–196, Florence, Italy. Association for Computational Linguistics.

AT McCray, A Burgun, and O Bodenreider. 2001. Aggregating umls semantic types for reducing conceptual complexity. *Proceedings of Medinfo*, 10(pt 1):216–20.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with {umls} concepts. In *Automated Knowledge Base Construction (AKBC)*.

Ishani Mondal, Sukannya Purkayastha, Sudeshna Sarkar, Pawan Goyal, Jitesh Pillai, Amitava Bhattacharyya, and Mahanandeeshwar Gattu. 2019. Medical entity linking using triplet network. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 95–100, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Isar Nejadgholi, Kathleen C. Fraser, Berry De Bruijn, Muqun Li, Astha LaPlante, and Khaldoun Zine El Abidine. 2019. Recognizing UMLS semantic types with deep learning. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 157–167, Hong Kong. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *arXiv e-prints*, page arXiv:1902.07669.

Denis Newman-Griffis, Albert M Lai, and Eric Fosler-Lussier. 2018. Jointly embedding entities and text with distant supervision. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 195–206, Melbourne, Australia. Association for Computational Linguistics.

John David Osborne, Binod Gyawali, and Thamar Solorio. 2013. Evaluation of YTEX and MetaMap for clinical concept recognition. *CEUR Workshop Proceedings*, 1179.

Roderic DM Page. 2011. Linking ncbi to wikipedia: a wiki-based approach. *PLoS currents*, 3.

Ahmad Pesaranghader, Stan Matwin, Marina Sokolova, and Ali Pesaranghader. 2019. deepBioWSD: effective deep neural word sense disambiguation of biomedical text data. *Journal of the American Medical Informatics Association*, 26(5):438–446.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. SemEval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62, Dublin, Ireland. Association for Computational Linguistics.

Matthew Purver and Stuart Battersby. 2012. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 482–491. Association for Computational Linguistics.

Priya Radhakrishnan, Partha Talukdar, and Vasudeva Varma. 2018. ELDEN: Improved entity linking using densified knowledge graphs. 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 1844–1853, New Orleans, Louisiana. Association for Computational Linguistics.

Jonathan Raphael Raiman and Olivier Michel Raiman. 2018. Deeptype: multilingual entity linking by neural type system evolution. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Ruth Reátegui and Sylvie Ratté. 2018. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Medical Informatics and Decision Making*, 18(3):74.

Kevin Reschke, Martin Jankowiak, Mihai Surdeanu, Christopher D Manning, and Daniel Jurafsky. 2014. Event extraction using distant supervision. In *Language Resources and Evaluation Conference (LREC)*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Mohammed Saeed, Christine Lieu, Greg Raber, and Roger G Mark. 2002. Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring. In *Computers in cardiology*, pages 641–644. IEEE.

Guergana K. Savova, Anni R. Coden, Igor L. Sominsky, Rie Johnson, Philip V. Ogren, Piet C. de Groen, and Christopher G. Chute. 2008. Word sense disambiguation across two domains: Biomedical literature and clinical notes. *Journal of Biomedical Informatics*, 41(6):1088 – 1100.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. 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.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2018. CLAMP a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336.

Audrey Tong, Lukas Diduch, Jonathan Fiscus, Yasaman Haghpanah, Shudong Huang, David Joy, Kay Peterson, and Ian Soboroff. 2018. Overview of the nist 2016 lorehlt evaluation. *Machine Translation*, 32(12):1130.

Debbie A Travers and Stephanie W Haas. 2006. Unified Medical Language System Coverage of Emergency-medicine Chief Complaints. *Academic Emergency Medicine*, 13(12):1319–1323.

Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2018. Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics*, 84:93–102.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. RESIDE: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Chang Wang and Sridhar Mahadevan. 2011. Heterogeneous domain adaptation using manifold alignment. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI11, page 15411546. AAAI Press.

M Weeber, J G Mork, and A R Aronson. 2001. Developing a test collection for biomedical word sense disambiguation. *Proceedings. AMIA Symposium*, pages 746–750.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv e-prints*, page arXiv:1910.03771.

Junshuang Wu, Richong Zhang, Yongyi Mao, Hongyu Guo, Masoumeh Soflaei, and Jinpeng Huai. 2020. Dynamic graph convolutional networks for entity linking. In *Proceedings of The Web Conference 2020*, pages 1149–1159.

Han Xiao. 2018. bert-as-service. `https://github.com/hanxiao/bert-as-service`.

Yadollah Yaghoobzadeh, Heike Adel, and Hinrich Schütze. 2017. Noise mitigation for neural entity typing and relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1183–1194, Valencia, Spain. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Sendong Zhao, Ting Liu, Sicheng Zhao, and Fei Wang. 2019. A Neural Multi-Task Learning Framework to Jointly Model Medical Named Entity Recognition and Normalization. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pages 817–824. AAAI Press.

Joey Tianyi Zhou, Ivor W.Tsang, Sinno Jialin Pan, and Mingkui Tan. 2014. Heterogeneous Domain Adaptation for Multiple Classes. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 1095–1103, Reykjavik, Iceland. PMLR.

A Datasets Type Coverage and Description

In this section, we present the type coverage of different medical entity linking datasets. For each dataset, we report the number of annotated mentions across all semantic types. The results are reported in Table 5. Overall, we find that our proposed datasets: WIKIMED and PUBMEDDS provide much better coverage of semantic types than all the existing datasets. Below we present as short description of the datasets used for the experiments in this paper:

- **NCBI:** The NCBI-disease corpus of (Doğan et al., 2014) consists of 793 PubMed abstracts annotated with disease mentions and their corresponding concepts in the MEDIC vocabulary (Davis et al., 2012).
- **Bio CDR:** The CDR corpus (Li et al., 2016) consists of 1,500 PubMed abstracts annotated with mentions of chemicals, diseases, and relations between them. These mentions were normalized to their unique concept identifiers, using MeSH as the controlled vocabulary.
- **ShARe:** The ShARe corpus data of (Pradhan et al., 2014) comprises 431 anonymized clincial notes, obtained from the MIMIC II clinical dataset (Saeed et al., 2002) and annotated with disorder mentions.
- **MedMentions:** The MedMentions data of (Mohan and Li, 2019) consists of 4,392 PubMed abstracts annotated with several biomedical mentions. Each mention is labeled with a unique concept identifier and a semantic type using the UMLS as the target ontology.

B Medical Entity Linkers

In this section, we present a short description of the different medical entity linking systems used in this work for probing the impact of incorporating a type-based candidate concepts filtering step.

- **MetaMap** (Aronson and Lang, 2010) leverages a knowledge-intensive approach based on symbolic NLP and linguistic techniques to map biomedical mentions in text to UMLS concepts.
- **cTAKES** (Savova et al., 2010) uses a terminology-agnostic dictionary look-up algorithm for mapping named entities to UMLS concepts. We utilize the Clinical Pipeline of cTAKES augmented with LVG Annotator[3].
- **MetaMapLite** (Demner-Fushman et al., 2017) re-implements the basic functionalities of MetaMap with an additional emphasis on real-time processing and competitive performance.
- **QuickUMLS** (Soldaini and Goharian, 2016) is a fast, unsupervised algorithm that leverages approximate, dictionary-matching techniques for mapping biomedical entities in text.
- **ScispaCy** (Neumann et al., 2019) builds upon the robust spaCy library (Honnibal and Montani, 2017) for several biomedical and scientific text-processing applications such as parsing, named entity recognition, and entity linking.

C Discussion

Combining MEDTYPE with Biomedical WSD methods Disambiguating the candidate concepts produced by medical entity extraction pipelines has been a long-standing area of research, with several tools developed to integrate with existing pipelines. The YTEX suite of algorithms (Garla et al., 2011; Garla and Brandt, 2012) extends both MetaMap and cTAKES with a disambiguation module that helps to reduce noise considerably, although Osborne et al. (2013) found that it often over-filtered correct concepts. These methods can be combined

[3] https://cwiki.apache.org/confluence/display/cTAKES/cTAKES+4.0+-+LVG

with MEDTYPE to create a multi-stage filtering approach for disambiguation. MEDTYPE performs coarse filtering to a high-confidence set based on predicted type, a key step for narrowing down over-generated candidate sets in open-ended deep learning systems; disambiguation methods can then perform a fine-grained selection of the correct candidate to improve further entity linking performance. We highlight this as an important direction for future work on medical entity linking.

D Supplementary Results

D.1 Entity Linking with MEDTYPE

In this section, we present the results on entity linking with MEDTYPE using two other evaluation metrics: *Partial_mention_id_match* and *Only_entity_id_match* as described in Section 6.3. The results are reported in Table 6 and 7 respectively. Similar to Section 7.1, we found that MEDTYPE helps to improve medical entity linking across all the settings and is comparable to the coarse-grained oracle.

D.2 Type-wise Analyzing of Improvement from WIKIMED and PUBMEDDS

In this section, we present the type-wise results for demonstrating the benefit of utilizing our proposed datasets: WIKIMED and PUBMEDDS for medical entity linkers. The results are reported for all four dataset: NCBI (Table 8), Bio CDR (Table 9), ShARe (Table 10), and MedMentions (Table 11). Overall, we find that utilizing our proposed datasets help improve performance across most of the semantic types. The gains are more substantial for the types which have less coverage in the original training data.

D.3 Analysis of Candidate Set Size Reductions with MEDTYPE

Figure 5 expands the analyses presented in Figure 3 to show the detailed distribution of the candidate set sizes within the analyzed samples of MedMentions. ScispaCy (presented here as the best-performing entity linker on MedMentions) limits its output candidate set to 5 by default; however, all linkers used displayed similar behavior in our experiments.

E Implementation Details

We use pre-trained weights of *bert-base-cased* (Devlin et al., 2018) and BioBERT (Lee et al., 2019) for initializing BERT component of MEDTYPE. MEDTYPE is implemented using HuggingFace Transformers library (Wolf et al., 2019). For training, we utilize Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001. Along with the training code, we also provide a scalable implementation of MEDTYPE based on *bert-as-service* (Xiao, 2018) for processing multiple documents simultaneously. We use the default hyperparameters for all the entity linkers and components of MEDTYPE. A grid search over the validation split was performed for deciding a threshold for each semantic type from the range of (0.001, 1). The area under the Precision-Recall curve was used for choosing the best threshold. All training was performed on NVIDIA-GTX 1080Ti GPUs. Each training epoch of MEDTYPE takes from 5 mins to 2 days depending on the size of the dataset. In terms of number of parameters, MEDTYPE has around 110 million parameters (same as BERT-base model).

Categories	NCBI	Bio CDR	ShARe	MedMentions	WIKIMED	PUBMEDDS
Activities & Behaviors	4	7	1	12,249	554	2,725,161
Anatomy	3	29	4	19,098	14,366	10,688,138
Chemicals & Drugs	0	32,436	1	46,420	26,809	44,476,957
Concepts & Ideas	0	0	1	60,475	2,562	5,274,354
Devices	0	0	0	2,691	483	242,599
Disease or Syndrome	10,760	22,603	5,895	11,709	84,706	9,846,667
Disorders	664	1,853	997	3,575	8,635	1,115,186
Finding	749	2,220	500	15,666	9,285	1,778,023
Functional Concept	0	0	1	23,672	117	48,553
Genes & Molecular Sequences	20	0	0	5,582	446	281,662
Living Beings	0	43	7	31,691	919,694	21,339,662
Mental or Behavioral Dysfunction	293	3,657	410	2,463	19,196	2,353,547
Neoplastic Process	4,022	2,301	323	4,635	16,823	1,476,843
Objects	0	129	2	10,357	421	5,184,355
Occupations	0	0	0	1,443	1,156	654,604
Organic Chemical	0	90,428	1	10,258	17,330	50,248,085
Organizations	0	0	0	2,276	0	298,119
Pathologic Function	143	3,290	2,285	4,121	4,474	1,895,835
Pharmacologic Substance	0	90,872	1	11,935	24,878	50,696,769
Phenomena	4	163	2	7,210	317	1,722,873
Physiology	15	166	3	24,753	2,054	10,674,561
Procedures	5	73	4	37,616	4,008	7,471,434
Qualitative Concept	0	0	7	32,564	106	1,211,747
Sign or Symptom	211	9,844	2,687	1,809	4,212	3,750,734

Table 5: Category distribution of different datasets. Overall, we find that our proposed datasets: WIKIMED and PUBMEDDS give diverse coverage across all semantic types.

	NCBI	Bio CDR	ShARe	MedMentions	WIKIMED
MetaMap	9.0	20.7	8.1	39.8	71.9
Oracle (Fine)	+0.5	+0.3	+0.3	+6.9	+0.8
Oracle (Coarse)	+0.5	+0.3	+0.3	+6.1	+0.6
MEDTYPE	+0.5	+0.2	+0.2	+4.1	+0.5
cTAKES	21.6	42.1	18.7	18.3	11.4
Oracle (Fine)	+0.2	+0.1	+0.1	+0.2	+0.1
Oracle (Coarse)	+0.2	+0.1	+0.1	+0.2	+0.1
MEDTYPE	+0.2	+0.1	0.0	+0.1	+0.1
MetaMapLite	7.3	18.9	7.5	35.2	66.4
Oracle (Fine)	+1.6	+1.5	+1.5	+7.8	+2.0
Oracle (Coarse)	+1.6	+1.5	+1.5	+6.5	+1.7
MEDTYPE	+1.6	+1.4	+1.4	+4.3	+1.5
QuickUMLS	7.2	13.2	5.8	10.1	6.2
Oracle (Fine)	+0.9	+3.7	+3.1	+2.9	+2.0
Oracle (Coarse)	+0.9	+3.2	+3.1	+2.4	+1.8
MEDTYPE	+0.8	+3.1	+2.7	+1.5	+1.7
ScispaCy	12.7	24.4	10.2	40.6	48.0
Oracle (Fine)	+2.0	+2.0	+2.0	+9.4	+2.2
Oracle (Coarse)	+2.0	+2.0	+2.0	+7.8	+2.0
MEDTYPE	+1.9	+1.7	+1.8	+4.4	+1.7

Table 6: We report the F1-score for partial mention and entity identifier match (Section 6.3). For each method, the first row is its base performance and following rows indicates the change in scores with different entity disambiguation methods.

	NCBI	Bio CDR	ShARe	MedMentions	WikiMed
MetaMap	10.8	23.4	16.0	60.9	79.3
Oracle (Fine)	0.0	+0.1	+0.7	+3.1	+0.6
Oracle (Coarse)	0.0	+0.1	+0.7	+2.5	+0.4
MedType	0.0	0.0	+0.5	+1.0	+0.2
cTAKES	30.8	46.0	31.9	31.0	18.3
Oracle (Fine)	0.0	+0.1	+0.1	+0.1	0.0
Oracle (Coarse)	0.0	+0.1	+0.1	+0.1	0.0
MedType	0.0	+0.1	0.0	0.0	0.0
MetaMapLite	11.8	25.0	16.6	58.1	76.7
Oracle (Fine)	+0.7	+0.8	+1.2	+5.0	+1.4
Oracle (Coarse)	+0.7	+0.8	+1.2	+4.1	+1.2
MedType	+0.6	+0.7	+1.0	+2.1	+0.9
QuickUMLS	10.8	21.4	14.3	24.8	12.9
Oracle (Fine)	+0.8	+2.6	+2.4	+1.9	+1.7
Oracle (Coarse)	+0.8	+2.3	+2.4	+1.6	+1.3
MedType	+0.8	+2.1	+1.9	+0.7	+1.1
ScispaCy	14.2	27.6	18.8	59.6	70.3
Oracle (Fine)	+1.1	+1.4	+1.3	+4.5	+1.9
Oracle (Coarse)	+1.1	+1.4	+1.3	+3.7	+1.8
MedType	+0.8	+1.0	+0.7	+0.7	+1.3

Table 7: We report the F1-score for only entity identifier match. For each method, the first row is its base performance and following rows indicates the change in scores with different entity disambiguation methods.

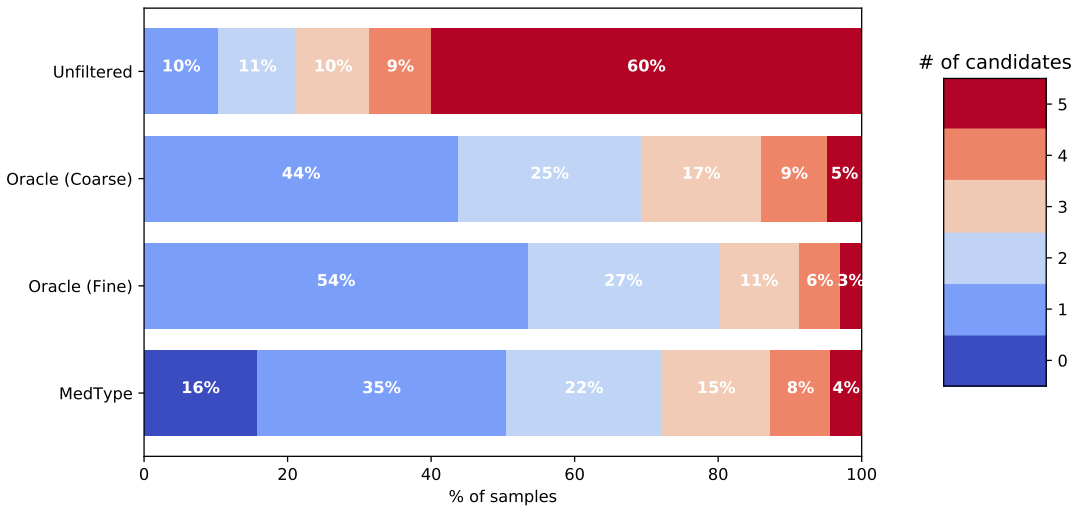


Figure 5: Distribution of candidate set sizes in MedMentions using ScispaCy, comparing unfiltered concepts to candidate sets filtered using oracle coarse-grained or fine-grained entity typing.

Categories	Training (T)	T ← WIKIMED	T ← PUBMEDDS	T ← Both	# Mentions
Activities & Behaviors	0	0	66.7	0	3
Anatomy	0	0	0	0	2
Chemicals & Drugs	0	0	0	0	0
Concepts & Ideas	0	0	0	0	0
Devices	0	0	0	0	0
Disease or Syndrome	91.2	95.1	96.7	96.1	4426
Disorders	25	58.5	68.5	69.7	417
Finding	0	30.8	45.8	43.9	158
Functional Concept	0	0	0	0	0
Genes & Molecular Sequences	0	0	0	0	14
Living Beings	0	0	0	0	0
Mental or Behavioral Dysfunction	0	76.5	80	76.5	73
Neoplastic Process	87.6	93.3	92.2	93.3	1224
Objects	0	0	0	0	0
Occupations	0	0	0	0	0
Organic Chemical	0	0	0	0	0
Organizations	0	0	0	0	0
Pathologic Function	0	75	50	66.7	93
Pharmacologic Substance	0	0	0	0	0
Phenomena	0	0	0	0	2
Physiology	0	0	0	0	9
Procedures	0	0	0	0	2
Qualitative Concept	0	0	0	0	0
Sign or Symptom	0	60	64.5	80	91

Table 8: Type-wise analysis of the impact on using WIKIMED and PUBMEDDS on NCBI dataset.

Categories	Training (T)	T ← WIKIMED	T ← PUBMEDDS	T ← Both	# Mentions
Activities & Behaviors	0	0	0	0	2
Anatomy	0	0	0	0	15
Chemicals & Drugs	66.3	81.2	88	88.9	6148
Concepts & Ideas	0	0	0	0	0
Devices	0	0	0	0	0
Disease or Syndrome	76.9	87.6	90.2	91.3	5047
Disorders	59	78.6	83.5	84.8	799
Finding	45.8	72.6	82.7	82.6	1054
Functional Concept	0	0	0	0	0
Genes & Molecular Sequences	0	0	0	0	0
Living Beings	0	0	44.4	40	11
Mental or Behavioral Dysfunction	46.6	83.4	86	87.7	707
Neoplastic Process	50.8	84.4	86.9	87.8	484
Objects	0	4.3	38.2	30.4	69
Occupations	0	0	0	0	0
Organic Chemical	81.2	89	92	92.2	8198
Organizations	0	0	0	0	0
Pathologic Function	60.6	81.6	87.3	89	1433
Pharmacologic Substance	82.3	88.1	90.6	90.7	8934
Phenomena	4.7	72.5	80.9	88.3	121
Physiology	0	39.4	64.3	57.1	107
Procedures	0	0	24.2	0	47
Qualitative Concept	0	0	0	0	0
Sign or Symptom	61.9	85.8	89.8	90.1	1379

Table 9: Type-wise analysis of the impact on using WIKIMED and PUBMEDDS on Bio CDR dataset.

Categories	Training (T)	T ← WIKIMED	T ← PUBMEDDS	T ← Both	# Mentions
Activities & Behaviors	0	0	0	0	1
Anatomy	0	0	0	0	1
Chemicals & Drugs	0	0	0	0	1
Concepts & Ideas	0	0	0	0	0
Devices	0	0	0	0	0
Disease or Syndrome	72.3	86.7	88.9	86.7	3457
Disorders	38.5	64.5	73.5	64.5	596
Finding	42.9	74.5	77.1	74.5	315
Functional Concept	0	0	0	0	0
Genes & Molecular Sequences	0	0	0	0	0
Living Beings	0	0	0	0	1
Mental or Behavioral Dysfunction	0	76.5	78.9	76.5	240
Neoplastic Process	0	81.4	74.1	81.4	182
Objects	0	0	0	0	0
Occupations	0	0	0	0	0
Organic Chemical	0	0	0	0	0
Organizations	0	0	0	0	0
Pathologic Function	54.8	80.2	85.6	80.2	1236
Pharmacologic Substance	0	0	0	0	1
Phenomena	0	0	0	0	0
Physiology	0	0	0	0	0
Procedures	0	0	0	0	2
Qualitative Concept	0	0	0	0	7
Sign or Symptom	78.5	91.2	93.2	91.2	1397

Table 10: Type-wise analysis of the impact on using WIKIMED and PUBMEDDS on ShARe dataset.

Categories	Training (T)	T ← WIKIMED	T ← PUBMEDDS	T ← Both	# Mentions
Activities & Behaviors	67	67.3	71.3	71.2	9597
Anatomy	69.1	71	78.4	78.9	15765
Chemicals & Drugs	67.9	67.6	74.3	75.2	37090
Concepts & Ideas	74.2	75.5	77.9	78.4	48231
Devices	31.1	33	40.1	37.6	2211
Disease or Syndrome	66.8	69.8	76.7	78	9475
Disorders	53.9	56.2	60.6	62	2868
Finding	49.1	52.7	55.3	56.3	12582
Functional Concept	73.6	74.5	75.4	77.1	18989
Genes & Molecular Sequences	56	57.3	60.3	60.7	4447
Living Beings	80.5	81.6	85.1	85.1	25216
Mental or Behavioral Dysfunction	70.5	70.6	76.2	74.7	1867
Neoplastic Process	71.4	73.4	78.3	78.8	3726
Objects	60.7	61.9	66.8	67.1	8161
Occupations	37.5	34.5	50.6	46.9	1094
Organic Chemical	56.6	60	72.6	72.2	8143
Organizations	54.1	55.2	62.1	63.1	1868
Pathologic Function	61.6	63.9	67.1	67.1	3384
Pharmacologic Substance	50.6	52	63.7	62.4	9659
Phenomena	45.1	46	55.1	55.9	5885
Physiology	66.9	68.4	70.5	72.8	19750
Procedures	68.8	69.5	74.1	74.9	30113
Qualitative Concept	79.9	80.9	82.2	82.7	25961
Sign or Symptom	67.9	69.8	72.6	74.1	1468

Table 11: Type-wise analysis of the impact on using WIKIMED and PUBMEDDS on MedMentions dataset.