Data and text mining

Creating a scalable deep learning based Named Entity Recognition Model for biomedical textual data by repurposing BioSample free-text annotations

Brian Y Tsui[1], Shamim Mollah[2], Dylan Skola[1], Michelle Dow[1], Chun-Nan Hsu[1] and Hannah Carter[1*]

[1]Department of Medicine, University of California San Diego, 9500 Gilman

[2]Department of Bioengineering, University of California San Diego, 9500 Gilman

*To whom correspondence should be addressed.

Abstract

Motivation:
Extraction of biomedical knowledge from unstructured text poses a great challenge in the biomedical field. Named entity recognition (NER) promises to improve information extraction and retrieval. However, existing approaches require manual annotation of large training text corpora, which is laborious and time-consuming. To address this problem we adopted deep learning technique that repurposes the 43,900,000 Entity- free-text pairs available in metadata associated with the NCBI BioSample archive to train a scalable NER model. This NER model can assist in biospecimen metadata annotation by extracting named-entities from user-supplied free-text descriptions.

Results: We evaluated our model against two validation sets, namely data sets consisting of short-phrases and long sentences. We achieved an accuracy of 93.29% and 93.40% in the short-phrase validation set and long sentence validation set respectively.

Availability: All the analyses, pre-trained model, environments, and Jupyter notebooks pertaining to this manuscript are available on Github: https://github.com/brianyiktaktsui/DEEP_NLP.

Contact: hkcarter@ucsd.edu

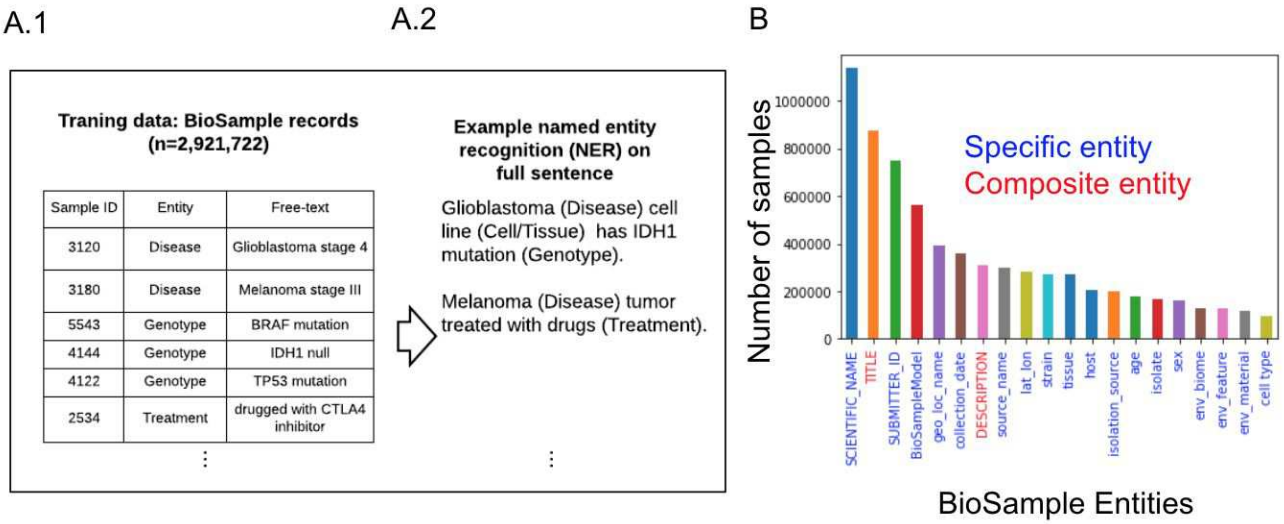


Fig1. Repurposing public biospecimen data for NER training (**A**) Depiction of training NER model using pre-annotated Entity- free-text pairs available from public biospecimen annotation data (BioSamples) from NCBI (**A.1**) Example of Entity- free-text pairs from BioSamples. In this example, the free-text phrase "Glioblastoma stage 4 system" is a "Disease" entity. (**A.2**) Expected results of an NER model recognizing biomedical concepts from sentences. (**B**) Histogram of the 30 most frequently used entities (x-axis) available in the current set of BioSamples. These atomic named entities (blue labels) can be used to extract concepts from composite entities TITLE and DESCRIPTION (red labels).

1 Introduction

Named Entity Recognition (NER) is an important task in the biomedical Natural Language Processing (NLP) domain. For example, NER can extract entities such as diseases, species, treatments, genes or geographical locations from biomedical free-text. A well constructed NER allows efficient document categorization and thus improve document retrieval accuracy.

Using NLP to automate the extraction of entity labels from biomedical textual data continues to be an area of active research. Biomedical NLP engines like MetaMap (Aronson, 2001), cTakes (Savova *et al.*, 2010) and TaggerOne (Leaman and Lu, 2016) have been successfully applied in biomedical research settings for automated information extraction by using a mix of rule-based algorithms and corpus building, where corpus is a set of free-texts annotated by curators for text model generation. However, the body of biomedical knowledge is constantly increasing in size as well as language complexity (Huang and Lu, 2016). This trend means that new biomedical text will be composed using words, sentence structures or entity types that were unseen by the model at the time of construction. Thus, these approaches are difficult to adapt to such changes as corpus generation is often limited by the high financial cost of hiring expert curators.

Another approach that has been adopted that involves extracting a set of predefined terms from free-text to increase the relevancy of retrieved data (Galeota and Pelizzola, 2017; Barrett *et al.*, 2013; Shah *et al.*, 2009), where the terms and their relationships are often curated in the form of an ontology. However, the maintenance of coherency in an ontology is often difficult and also requires manual curation. In summary, existing approaches for NER do not scale well because of their dependence on manual curation.

In recent years, deep learning techniques have replaced much of the pre-processing steps required to extract knowledge from free-text. Some of the techniques include 1) word embedding to represent the concepts of words in dense format (Mikolov *et al.*, 2013), and 2) recurrent neural network (RNN) which utilizes Long Short-Term Memory (LSTM) cells to capture the dependency of words (Sutskever *et al.*, 2014). Recently, studies have shown that deep learning was capable of a generating an NER with high accuracy in biomedical text (Zhu *et al.*, 2018; Wu *et al.*, 2017). However, their approaches require a large training corpus.

Here, we utilized the millions of BioSample (Barrett *et al.*, 2012) metadata annotations hosted by NCBI to train a deep learning based NER model. The BioSample metadata is natively encoded as Entity-free-text pairs and contains over 1,000,000 sample annotations spanning over 100,000 studies, making it a great resource for training deep-learning-based NER models. The BioSamples are comprised of the NCBI primary archives, including the Sequence Read Archive (SRA) (Kodama *et al.*, 2012) and the database of Genotypes and Phenotypes (dbGAP) (Mailman *et al.*, 2007).

Given the community driven nature of BioSample reporting, we believe that the NER model can be kept up to date with the evolving biomedical vocabulary by simply retraining with the latest BioSamples. Therefore, we evaluated the potential of repurposing the vast amount of BioSample annotations for training an NER model. First, we showed that word embedding can be used to increase the sample coverage for model training. We then adopted deep learning techniques to repurpose these BioSamples to train a scalable NER model (Fig. 1A), by first generating a short phrase classification model with an accuracy of 93.29%. Subsequently, we constructed an NER model for complete sentences, by utilizing the scores emitted by the short phrase classification and N-gram model (Brown *et al.*, 1992). We achieved an accuracy of 93.4% for this NER model. We also compared this NER model with MetaMap (Aronson, 2001).

2 Method

2.1 BioSample data landscape

The BioSample annotation records were downloaded from the NCBI FTP website (ftp.ncbi.nlm.nih.gov/sra/reports/Metadata/) on May 15, 2018. The XML-encoded sample files were parsed into a python pickle object to allow simple data loading. Only ASCII characters from the BioSample are retained and python package spaCy was used for word tokenization with default parameters. We retrieved 2,921,722 BioSample records (SRA BioSample Symbol: SRS), spanning over 106,110 studies. The resulting python pickle object for SRS was 1.4GB. Each BioSample record is associated with a single biospecimen and has the metadata encoded in entity- (attribute) - free-text pairs. An example record is shown in Figure 1A.1. Details for BioSample annotation structure can be found at https://www.ncbi.nlm.nih.gov/biosample/docs/attributes/ and https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/.

The BioSample entity fields are diverse (n=19,996) as they are designed to capture the experimental conditions from over 100,000 studies. The retrieved biospecimen records consist of 43,907,007 Entity-free-text pairs. The number of unique text entries associated with each entity scales with the number of SRS records (Supplementary Fig.1),

suggesting BioSample annotation data can capture annotation diversity. The 145 BioSample entities are associated with over 10,000 biospecimen records and the top 30 most frequently used entities are listed in Figure 1B, suggesting a large volume of annotation data is available from BioSample for training a deep learning based phrase classification model.

DESCRIPTION and TITLE are among the top 10 most frequently used entries in BioSample (Fig. 1B). These composite entity fields have a mean length of 46.10 characters, 3.19-fold longer (Fig. 2) than the rest of the free-text fields (mean=14.46 characters, 95% confidence interval: 1.0 - 65.0 characters, Supplementary Fig. 2), suggesting that they could be made up of specific experimental entities like source name, age, sex, etc. This begs the question of whether we can reclassify those annotations using an NER trained by the more atomic entities.

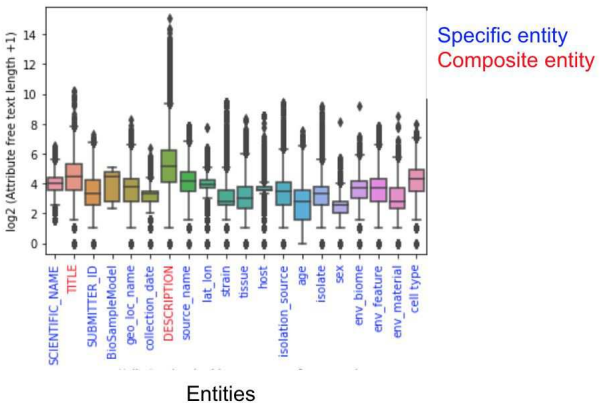


Fig. 2 Text length distribution of top entities

2.2 Vectorization of biomedical words to expand data coverage in NER model training

2.2.1 Word embeddings retrieval

Traditionally, words have been represented by a sparse hot-one encoding scheme in NLP. However, this trend changed when word embedding was introduced (Mikolov *et al.*, 2013). In principle, the semantic similarity of words should be representable by geometric distances between trained word embeddings. This approach has the advantage of not requiring the laborious process of pre-defining the meaning of words by hard-coding semantics for each word or generating rules for part-of-speech tagging. For this project, we used a published biomedical word embedding model (Chiu et al. 2016) that was trained on the entire PubMed, PMC and Wikipedia text corpora, with a total of 5,443,656 word vectors where each word is represented by 200 features.

2.2.2 Evaluation of semantic similarities at word, sentence, and entity resolution using word embedding

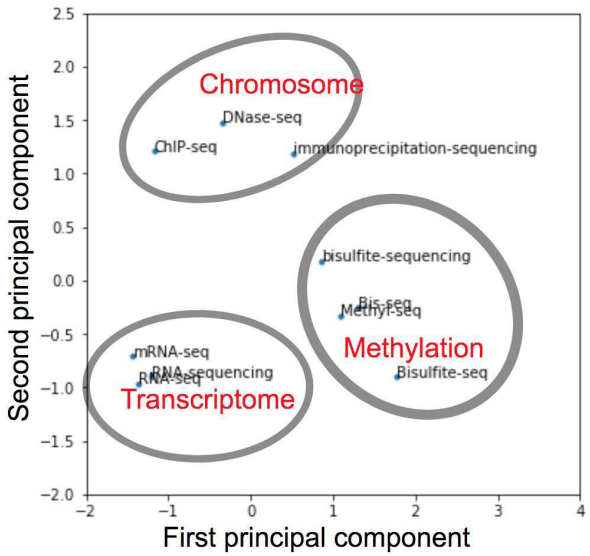


Fig. 3 Word embeddings group related terms in PCA space

The first validation of the utility of word embedding for this task was in the automatic identification of semantically similar words. The high number of word vectors enabled capture of word variations. For example, the word embedding model grouped related sequencing keywords in the PCA (Principal Component Analysis) space (Fig. 3; variance explained: PC0: 34.2%; PC1: 20.4%).

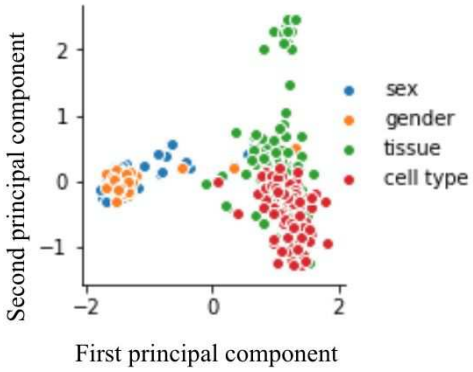


Fig. 4 Native sentence embeddings recover reasonable grouping

The second utility of word embedding we found was in combining semantically similar entities to increase BioSample coverage for model building and validation, as it has been recognized that there is a lack of standardization in metadata annotation (Brazma *et al.*, 2001). To summarize and represent a free-text sentence, a sentence embedding vector can be generated by taking the feature-wise mean of the embedding vectors of the words with the free-text. (Iyyer *et al.*, 2015) For example, sex is synonymous with gender and tissue is synonymous with cell type. When we randomly drew 100 biospecimen records with *sex*, *gender*, *tissue* and *cell type* BioSample entities, the sentence embedding vectors were able to group the free-texts according to their semantics (Fig. 4). To summarize at the entity level, we further reduced the sentence level embeddings to entity level embeddings using the same method. The entity embedding vectors of the 30 most commonly used BioSample entities cluster by cosine similarity, which further shows

higher embedding similarities for the more semantically similar entities (Supplementary Fig. 3).

2.2.3 Expansion of NER training data coverage using entity embedding similarities

The entities selected for the NER model are the commonly used metadata attributes including Species, Genotype, Disease state, Cell type/tissue, Geographical Location, and Treatment/Conditions. To identify samples in each class from BioSample archive, we first incorporated the corresponding official BioSample attributes: SCIENTIFIC_NAME, genotype, disease, cell type, geo_loc_name, treatment. Then, to increase the sample coverage for each entity class in NER training, we identified all the entities with a high cosine similarity (>0.8) with the corresponding official BioSample entity, and all those retrieved entities deemed synonymous from the label names (Supplementary Table 1), suggesting that all the entities in the BioSample with high embedding similarities can be potentially merged to reduce the complexity of the database and increase sample size for each equivalent class. For example, the sample coverage of genotype entity increased by 62.65% after merging the BioSample entities with high embedding similarities.

2.3 NER model training and validation

An NER model takes in free-text of any length and extracts the biomedical entities in phrases. Traditional entity recognition models require the corpus to be in complete sentences with entity labels annotated in order for the model to segment and identify phrase boundaries, as suggested by the popular corpus format ConNLL (Tjong Kim Sang 2002). However, this approach is not useful for the BioSample metadata which is encoded as Entity- free-text pairs. Therefore, we hypothesized that we could first train a short phrase entity classification model and then use an n-gram approach to extract named entity segments in longer sentences by identifying n-grams with high emission probability (Fig. 5).

2.3.1 Short phrase entity classification model construction and validation

In order to train the short phrase entity classification model, we retained only text comprising 2 to 7 words. The upper bound of 7 words corresponds to the 95th percentile of the distribution of phrase lengths in a highly curated general purpose medical vocabulary called the NCI Thesaurus (Sioutos *et al.*, 2007/2). This filtering also has the advantage of limiting the number of parameters in the Recurrent Neural Network (RNN) model. To build and validate the deep learning based NLP model for short phrase entity classification, we first randomly split the training and validation cohorts in a 4:1 ratio by studies, to maximize training-validation set independency. The deep NLP model is trained based on a bidirectional RNN architecture with LSTM cells (biLSTM) (Graves and Schmidhuber, 2005).

For both the training and testing data, we first converted the biospecimen free-text to a sentence by word ID matrix, where each row in the matrix consists of a sequence list of word embedding IDs. Then, the model was constructed and trained with the following layers:

1. An embedding layer to convert word embedding IDs to word vectors.
2. A bidirectional layer with a total of 64 hidden units, and a dropout rate of 0.5.
3. A dense, fully connected layer with logistic activation function for outputting the probability score used for classification. The number of neurons in this layer is the same as the number of entities.

The Adam optimizer (Kingma and Ba, 2014) with categorical cross-entropy as loss function was used to compile the deep learning model. This biLSTM model was constructed and trained using keras v2.7.1 with tensorflow v1.8.0 backend on a single machine with 48 physical cores with four Intel(R) Xeon(R) CPU E5-4650 v3 @ 2.10GHz, with a learning rate of 0.001 and batch size of 100.

2.3.2 Long sentence NER model construction and validation

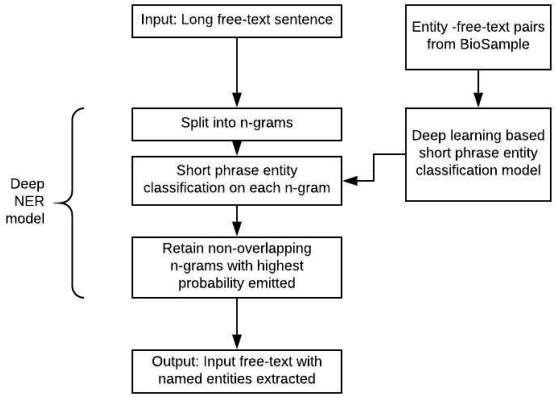


Fig. 5 Depiction of NER model

Each free-text input is segmented into sentences using commonly used separators like ':' , '.' , ',', In total, 179 stop words from the python package NLTK were removed from the input sentence. For each sentence, the non-alphanumeric characters were removed. To generate entity scores for each n-gram, the algorithm scans for entities among all phrases that match phrase lengths specified during model construction, by applying the short phrase classification model on each n-gram with n ranges from 2 to 7 in the sentence. Only n-grams with at least 2 tokens matching with word embedding IDs are retained. The tokens were then converted to word embedding IDs and became input to the predictive function of the short phrase entity classification model. Therefore, each n-gram will be associated with a vector of entity scores emitted by the logistic activation function from the last layer of the biLSTM, ranging from 0.0 to 1.0.

We also estimated the baseline neural net emission entity scores by executing the prediction function without any input. All the n-grams from emission vectors with absolute sum difference < 0.01 with baseline emission are zeroed. For overlapping n-grams that have the same entity, only the n-gram with the highest probability score will be retained.

2.4 Data generation for validation and method comparison

2.4.1 Manual curation set

To compare our NER model against manual curation, we manually curated 185 sentences selected from 100 randomly drawn BioSample records with a DESCRIPTION entity using Dataturk (https://dataturks.com/). For each sentence, we highlighted the token segments that described the entities selected in model construction. We then downloaded the annotated data in JSON format and compared it against the prediction results.

2.4.2 MetaMap set

We compared the performance of our NER model against MetaMap from NLM, which has been adopted by GIANT (Greene *et al.*, 2015) for automating biospecimen annotation extraction. We used the online MetaMap service (URL: https://ii.nlm.nih.gov/Batch/UTS_Required/MetaMap.shtml) to extract the terms from the validation cohort. In order to match the UMLS semantic types generated from MetaMap to our entities, we mapped the UMLS semantic types to UMLS groups using the table from the MetaMap website: https://MetaMap.nlm.nih.gov/Docs/SemGroups_2013.txt.

3 Results

3.1 Short phrase entity classification model performance

The short phrase entity recognition model was able to classify 93.29% percent of the data correctly in the validation cohort. All the entities were able to achieve an ROC-AUC (Fig. 6a) and an accuracy (Fig. 6b) over 90% in the validation cohorts, suggesting a good sensitivity and specificity. Next, we also observed a high F1 score of 92.14% across all the entities, suggesting our the model also has high precision and recall over all the classes.

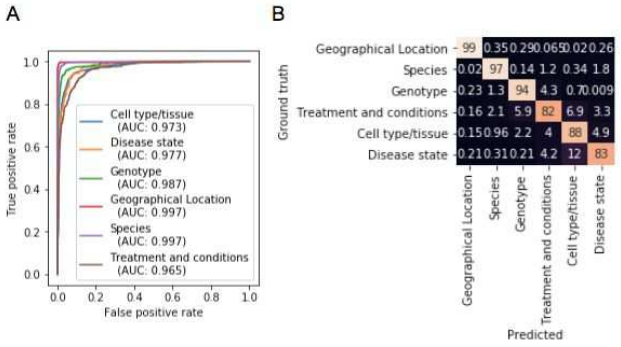


Fig. 6 Short phrase classification model performance (A) ROC-AUC of entities (B) Contingency table quantifying percentage of entities predicted correctly in validation cohort.

Also, the accuracy of the short phrase entity recognition model improved when trained with more data (Fig. 7), confirming the utility of using the ever growing source of community-contributed BioSample data to train a more accurate NER model.

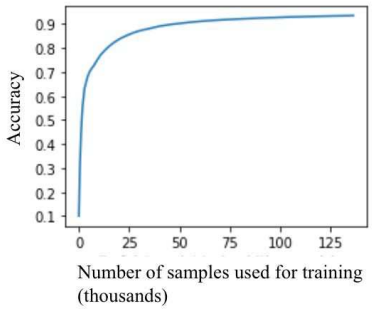


Fig. 7 Volume of data scale with training performance

3.2 NER performance in long sentence

Since BioSample metadata is in the format of entity- free-text pairs as opposed to a fully annotated corpus, we evaluated the possibility of using an n-gram approach to segment the text. For each n-gram, we assign the entity with the highest short phrase classification score.

We compared the performance of the model against manual curation and the existing tool MetaMap (Aronson, 2001) from NLM. We manually curated 100 free-text entries from samples with the DESCRIPTION field, with a total of 185 sentences where 139 sentences had at least one entity extracted in curation. The model yielded an accuracy of 0.934 in entity memberships recovery for each sentence. We then compared the performance of the model against MetaMap. For the task of biological entity recognition, our NER model had superior performance in terms of precision, sensitivity and specificity and F1-score (Table 1).

The poor precision and recall in MetaMap are likely due to the high level of multi-mapping (perplexity) among the terms in the output. For example, the phrase "liver hepatoblastoma" from validation annotation record SRS1098269 is mapped only to the Disease entity in our data, while MetaMap assigned mappings to Liver (Tissue) and Hepatoblastoma (Disease) independently, and also mapped the word "liver" to less relevant entities like Food. In addition, our model was able to extract phrases like "germline sp1" correctly as Genotype, while MetaMap mapped "germline" as a cell type "Germ Line" and did not recognize "sp1" as Genotype related. Also, "sc 197" (ERS396144) is an antibody was first developed in 2010. MetaMap trained using the MedTag (Smith *et al.*, 2005) corpus that was last updated in 2005 according to the file transfer program (FTP) timestamp, thus explaining why it could not capture this entity that was described in recent years. Finally, instead of recognizing "MRG15 null" (DRS033379) as a single entity Genotype like our model, MetaMap recognized the word "MRG15" as a Gene and null as an Qualitative concept.

	Deep bio NER	MetaMap
Sensitivity/recall	93.21%	73.86%
Specificity	94.09%	68.69%
Precision	82.68%	47.72%
F1-score	87.63%	57.98%

Table 1: NER performance comparison with MetaMap

4 Discussion

Here, we trained a deep learning based NER model by repurposing the vast amount of BioSample metadata in NCBI, available as entity-free-text pairs. We first showed that the word embeddings were able to group the free-text annotations at various resolutions, i.e. word, sentence and entity level. This allowed us to utilize entity embeddings to merge synonymous entity labels to increase sample coverage for short phrase entity classifier training. The short phrase entity classification model was effectively able to extract entities from phrases in the validation set. We then used this model to extract named entities from long sentences. We were able to achieve a higher accuracy using our NER model when compared with existing method MetaMap (Aronson, 2001).

Existing methods require a large annotated data corpus that is difficult to come by. For example, existing biomedical NER has relied on laborious corpus curation and annotation of the entities in each sentence. Expert curation is costly in terms of both time and monetary expenses, and may suffer from curation bias. For example, MetaMap (Aronson, 2001) relied on the MedTag (Smith *et al.*, 2005) corpus from NLM, which has only 15,000 sentences in total and was last updated in 2005 based on the FTP timestamp, which means that this corpus may not provide coverage for new domain knowledge.

In contrast, our approach can incorporate new vocabulary and entity definitions without the need for human intervention. The word embeddings (Chiu et al. 2016) used here were generated by PubMed articles and Wikipedia, and entity-free-text pairs were generated by the biomedical research community. We then utilized neuron emission strengths from the trained NER for phrase segmentation. This entire process was free of any manual curation. As a result, this model can be readily extended in vocabulary and NER capability by simply incorporating more training examples or new embeddings as they become available.

Furthermore, our model has a concise code base which improves readability and extendability. For example, the code base in cTakes consists of 1,404 java files with a total of 234,388 lines of code, and the code base in MetaMap consist of 218 files with 260,586 lines of code, while our model can be fully specified in less than 200 lines of code. Our approach eliminates the need for stemming, part of speech tagging (POS) and dependency parsing. Thus, deep learning simplifies model construction and provides good performance for automated biological entity recognition.

We acknowledged that there are limitations to our approach. For example, there is no professional curator to ensure the correctness of the BioSample entries which may have contributed to lower accuracy of our NER prediction. In addition, we incorporated a limited number of

entities into the model. Further evaluation is merited to determine whether adopting a Convolution Neural Network (CNN), or adding Conditional Random Fields (Zheng *et al.*, 2015) or attention layers (Luong *et al.*, 2015) will further improve the model. Also, we did not include any numerical entities due to the lack of availability of annotations (only the age entity had more than 10,000 biospecimen annotations). In the future, we are also interested in better understanding the confounding relationships between the entities in the BioSample archive and developing algorithms to identify synonyms entities, possibly by using community finding algorithms like Clixo (Kramer *et al.*, 2014).

4 Acknowledgements

We thank all members of the Carter and Ideker lab for scientific feedback and comments.

5 Funding

This work was funded by NIH grants DP5-OD017937, RO1 CA220009 and a CIFAR fellowship to H.C.

Conflict of Interest: none declared.

6 References

Aronson,A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.*, 17–21.

Barrett,T. *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–63.

Barrett,T. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.

Brazma,A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.

Brown,P.F. *et al.* (1992) Class-based N-gram Models of Natural Language. *Comput. Linguist.*, **18**, 467–479.

Galeota,E. and Pelizzola,M. (2017) Ontology-based annotations and semantic relations in large-scale (epi)genomics data. *Brief. Bioinform.*, **18**, 403–412.

Graves,A. and Schmidhuber,J. (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.*, **18**, 602–610.

Greene,C.S. *et al.* (2015) Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.*, **47**, 569–576.

Huang,C.-C. and Lu,Z. (2016) Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief. Bioinform.*, **17**, 132–144.

Iyyer,M. *et al.* (2015) Deep unordered composition rivals syntactic methods for text classification. In, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*., pp. 1681–1691.

Kingma,D.P. and Ba,J. (2014) Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]*.

Kodama,Y. *et al.* (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–6.

Kramer,M. *et al.* (2014) Inferring gene ontologies from pairwise similarity data. *Bioinformatics*, **30**, i34–42.

Leaman,R. and Lu,Z. (2016) TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics*, **32**, 2839–2846.

Luong,M.-T. *et al.* (2015) Effective Approaches to Attention-based Neural Machine Translation. *arXiv [cs.CL]*.

Mailman,M.D. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, **39**, 1181–1186.

Mikolov,T. *et al.* (2013) Distributed Representations of Words and Phrases and their Compositionality. In, Burges,C.J.C. *et al.* (eds), *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., pp. 3111–3119.

Savova,G.K. *et al.* (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc.*, **17**, 507–513.

Shah,N.H. *et al.* (2009) Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinformatics*, **10 Suppl 2**, S1.

Sioutos,N. *et al.* (2007/2) NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.*, **40**, 30–43.

Smith,L.H. *et al.* (2005) MedTag: a collection of biomedical annotations. *Proceedings of the ACL*.

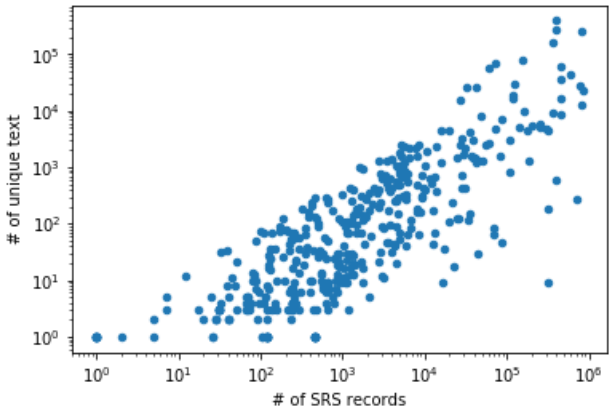
Sutskever,I. *et al.* (2014) Sequence to Sequence Learning with Neural Networks. In, Ghahramani,Z. *et al.* (eds), *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., pp. 3104–3112.

Wu,Y. *et al.* (2017) Clinical Named Entity Recognition Using Deep Learning Models. *AMIA Annu. Symp. Proc.*, **2017**, 1812–1819.

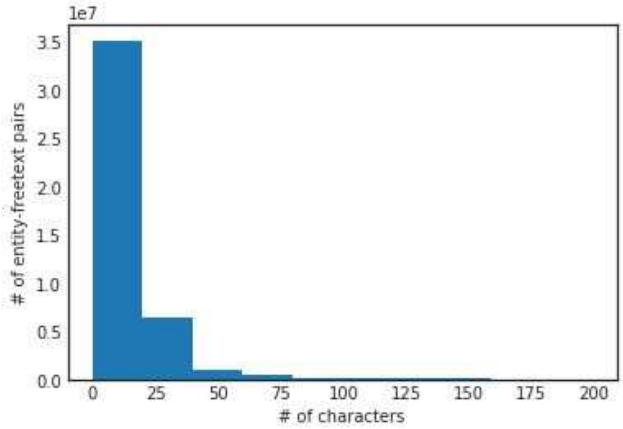
Zheng,S. *et al.* (2015) Conditional random fields as recurrent neural networks. In, *Proceedings of the IEEE international conference on computer vision*., pp. 1529–1537.

Zhu,Q. *et al.* (2018) GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics*, **34**, 1547–1554.

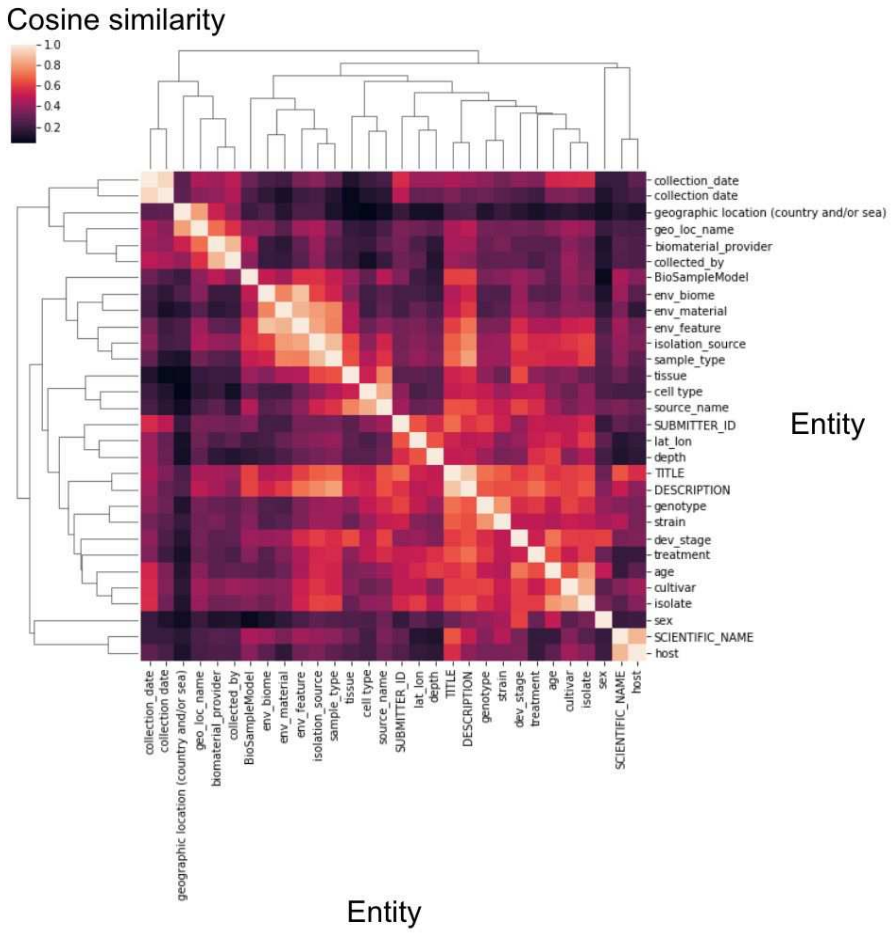
Supplementary Figures and Tables



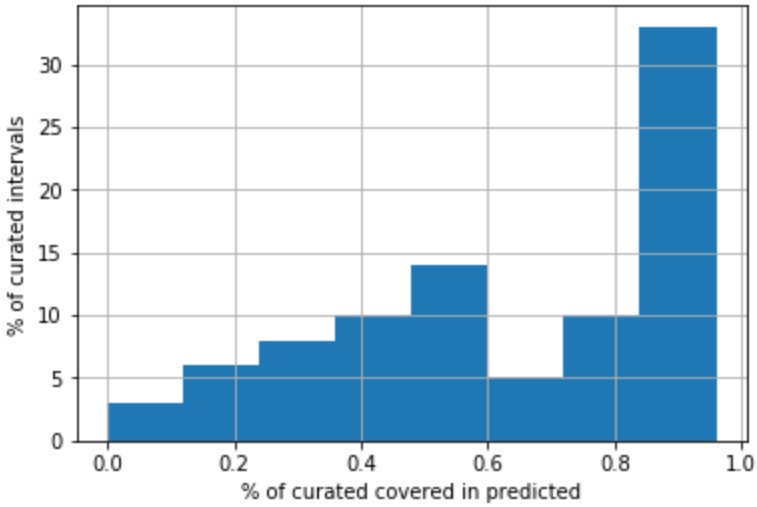
Supplementary Fig. 1 Number of unique free text annotations scales with the number of SRS records for each attribute.



Supplementary Fig. 2 Distribution of free text length (x-axis) over SRA entity- free-text pairs (y-axis)



Supplementary Fig. 3 Entity embedding recovers biospecimen entity similarities



Supplementary Fig. 4: Histogram of percentage of curated text overlap with NER predicted text (x-axis) over curated spans (y-axis).

Entity class name	BioSample entity name	Cosine similarity	# of samples per BioSample entity	# of samples (aggregated)	Fold increase
Species	SCIENTIFIC_NAME	1.00	1,136,856	1,539,081	1.35
	organism	0.98	29,037		
	Organism	0.90	2,894		
	host scientific name	0.86	9,909		
	Species	0.85	578		
	host	0.84	205,511		
	specific host	0.83	11,114		
	host_scientific_name	0.83	4,297		
	host organism	0.83	372		
	nat-host	0.82	1,516		
	specific_host	0.81	14,088		
Genotype	genotype	1.00	75,566	122,909	**1.63**
	genotype/variation	0.97	28,730		
	plant genotype	0.88	195		
	mutant	0.85	314		
	mutation	0.83	428		
	phenotype	0.82	13,892		
	host_genotype	0.80	3,784		
Disease state	disease	1.00	21,654	34,239	**1.58**
	tumor type	0.87	691		
	diagnosis	0.86	3,351		
	disease state	0.85	3,715		
	DiseaseState	0.83	173		
	cancer type	0.82	311		
	tumor	0.82	355		
	clinical history	0.82	280		
	disease status	0.82	2,077		
	cell description	0.80	1,632		
Cell type/tissue	cell type	1.00	94,819	417,924	**4.41**
	cell description	0.94	1,632		
	cell_type	0.93	18,162		
	source cell type	0.93	106		
	cell types	0.91	160		
	source_name	0.91	297,941		
	cell-type	0.89	286		
	CellType	0.88	342		
	cell subtype	0.88	1,220		

	biomaterial_type	0.86	182		
	progenitor cell type	0.86	169		
	tissue/cell type	0.85	783		
	DIFFERENTIATION_STAGE	0.84	170		
	cell	0.83	1,616		
	differentiation status	0.82	111		
	cell line source	0.81	225		
Geographical Location	geo_loc_name	1.00	389,901	509,997	1.31
	geographic location	0.96	13,640		
	geo loc name	0.94	704		
	geographic location (country and/or sea, region)	0.92	7,920		
	geographic location (country and/or sea,region)	0.87	12,833		
	country	0.86	24,413		
	geographic location (country and/or sea,region)	0.84	539		
	birth_location	0.83	3,182		
	geographic location (country and/or sea)	0.81	56,576		
	Geo_loc_name	0.80	289		
Treatment and conditions	treatment	1.00	73,041	90,516	1.24
	treated with	0.92	3,026		
	treatment protocol	0.91	583		
	drug treatment	0.89	977		
	agent	0.89	2,570		
	stimulated with	0.86	247		
	protocol	0.85	2,769		
	Treatment	0.84	3,591		
	sample group	0.82	1,176		
	cell treatment	0.82	231		
	experimental condition	0.82	187		
	culture conditions	0.81	391		
	culture condition	0.80	1,220		
	treatment group	0.80	405		
	sample treatment	0.80	102		

Supplementary Table 1: Entity embedding can recover synonymous BioSample entity labels which can increase the sample size in each entity class.