Automated ontology-based annotation of scientific literature using deep learning

Prashanti Manda
University of North Carolina at Greensboro
Greensboro, North Carolina 27412
p_manda@uncg.edu

Saed SayedAhmed
University of North Carolina at Greensboro
Greensboro, North Carolina 27412
smsayeda@uncg.edu

Somya D. Mohanty
University of North Carolina at Greensboro
Greensboro, North Carolina 27412
sdmohant@uncg.edu

CCS CONCEPTS

•Applied computing →Bioinformatics; •Information systems →Specialized information retrieval; *Information retrieval;*

KEYWORDS

Deep Learning, Automated Annotation, Named Entity Recognition, Ontologies

ACM Reference format:
Prashanti Manda, Saed SayedAhmed, and Somya D. Mohanty. 2020. Automated ontology-based annotation of scientific literature using deep learning. In *Proceedings of Semantic Big Data , Portland, OR, USA, June 14–19, 2020 (SBD'20),* 6 pages.
DOI: 10.1145/3391274.3393636

1 ABSTRACT

Representing scientific knowledge using ontologies enables data integration, consistent machine-readable data representation, and allows for large-scale computational analyses. Text mining approaches that can automatically process and annotate scientific literature with ontology concepts are necessary to keep up with the rapid pace of scientific publishing. Here, we present deep learning models (Gated Recurrent Units (GRU) and Long Short Term Memory (LSTM)) combined with different input encoding formats for automated Named Entity Recognition (NER) of ontology concepts from text. The Colorado Richly Annotated Full Text (CRAFT) gold standard corpus was used to train and test our models. Precision, Recall, F-1, and Jaccard semantic similarity were used to evaluate the performance of the models. We found that GRU-based models outperform LSTM models across all evaluation metrics. Surprisingly, considering the top two probabilistic predictions of the model for each instance instead of the top one resulted in a substantial increase in accuracy. Inclusion of ontology semantics via subsumption reasoning yielded modest performance improvement.

2 INTRODUCTION

The majority of scientific knowledge resides in the form of free text in scientific publications making it unavailable for large-scale computational inquiry [7]. Ontologies have been developed to promote consistent usage of terminology, data representation, and standardization. Ontologies are semantically rich data representation formats for precise and large-scale description of objects. Since their advent in 2003, biological ontologies have been widely used to represent data from scientific literature in a machine readable format that enables computational analyses. These analyses include studies of over represented functional gene ontology categories, semantic similarity of genes/phenotypes/diseases, etc. The majority of the currently available ontology annotations (descriptions of biological entities using ontology concepts) are generated manually by scientists who read literature and tag pieces of text with appropriate ontology concepts. The speed of manual curation cannot keep up with the rapid rate of scientific publishing creating a severe bottleneck that hampers knowledge generation and discovery.

Text mining and natural language processing techniques have been in development in response to this bottleneck. The goal of these approaches is to automatically process scientific literature and tag phrases of text with appropriate concepts from one or more ontologies. Text mining approaches for automated ontology annotation can be broadly classified into three categories: 1) Lexical/syntactic, 2) Machine learning, and 3) Deep learning [20].

Lexical approaches use lexical and semantic similarities between a piece of text and an ontology concept to annotate the text with the concept [20]. Other information sources in the ontology such as concept cross-references, definitions, and synonyms can also be used to perform string matching. These approaches have been shown to be challenging and error prone considering that some ontology concepts contain a large number of words which makes text matching difficult [20].

Machine learning based methods have used supervised learning techniques to train classifiers on known gold standard annotation corpora for identifying associations between text and ontology concepts. These methods are typically more successful as compared to lexical approaches since they are able to form generalized associations that aren't limited to lexical similarities. k-Nearest Neighbors has been a widely used method in this category [20]. Note that these methods depend on the availability of human curated data for training and testing.

Over the past few years, deep learning methods have been shown to have greater accuracy for text-based tasks [9, 11, 12, 22], in particular, named entity recognition of ontology concepts from text. Conventional machine learning methods can be limited in their ability to represent and encode large pieces of text thereby limiting their ability to make associations between text and ontology concepts [5], On the other hand, deep learning models use vector

SBD'20, Portland, OR, USA
 978-1-4503-7974-8/20/06…$15.00
DOI: 10.1145/3391274.3393636

representations that can encode notions of word dependence, context, word sequences leading to richer embeddings of the input data.

In this study, we present deep learning models for the task of automated annotation of ontology concepts from text. We explore the impact of various input encoding methods on the performance of the models. We use Gene Ontology [1] annotations, one of the most widely used ontologies in biology, from the CRAFT gold standard corpus [2] to train and test our models. Precision, Recall, F-1 score, and Jaccard semantic similarity are used to measure the accuracy of our models.

3 RELATED WORK

The application of deep learning methods to automatically annotate scientific literature with ontology concepts is relatively new. However, a number of promising studies have examined the potential of deep learning approaches on related applications. Deep learning tool kits such as word2vec have been applied to identify pharmaceutical properties from medical literature [15]. While not a direct application of ontology-powered annotation, results point to strategies for improvement that could enable more sophisticated applications. Similarly, deep neural networks have been applied to identify phenotypes (represented by ontologies) from exome or whole genome sequence data [4].

In direct applications of deep learning for recognizing ontology concepts, convolutional neural networks (CNN) combined with long short term memory models (LSTM) were used [21]. This work demonstrated the efficiency of deep learning methods to reduce the need of labeled training data while maintaining prediction accuracy. CNNs were also used for biomedical named entity recognition combined with n-gram character embeddings resulting in enhanced performance in a comparison with other deep learning models [25]. A comprehensive review of deep learning methods for named entity recognition can be found in [13] and a comparison of existing text mining tools in [3].

This study builds on our prior work [13] where we conducted an evaluation of deep learning architectures and evaluated the models on annotations from various ontologies extracted from 67 articles in the CRAFT gold standard corpus. The limitations of our prior work were that annotations were restricted to unigrams (annotations comprising of more than one word were excluded) and the models did not incorporate ontology hierarchy and semantics for prediction. Here, we make methodological improvements on those two fronts and present more sophisticated deep learning models for the task of ontology based named entity recognition from scientific literature.

4 METHODS

4.1 Dataset

The Colorado Richly Annotated Full-Text (CRAFT) corpus v3.0 containing ontology annotations for 97 articles was used for training and testing the models developed in this study [2]. The articles in this corpus are annotated with ontology concepts to include structural, coreference, and concept annotations.

4.2 Data Preprocessing

Annotations from the 97 articles in the CRAFT corpus were preprocessed as detailed in our previous study [13]. Ontology annotations were encoded using the *IOB* format [19]. Beginning words of an annotation are tagged with a *B* tag, while words inside an annotated phrase are tagged with an *I* tag. Words that are not part of any annotations are tagged with an *O* tag. Only Gene Ontology annotations were used in the experiments in this study since GO annotations account for the majority of annotations in the CRAFT corpus.

4.3 Deep learning models

In previous work, a set of eightmodels were tested at the task of ontology based NER. Results showed that GRUs and LSTMs outperformed RNNs and other models. Guided by those results, here, we developed models based on GRUs and LSTMs coupled with diferent input encodings such as GloVe, ELMo, etc.

4.3.1 Gated Recurrent Unit. Gated Recurrent Units (GRU) [6] consist of two gates - Update and Reset. These gates control the information to be held and passed to the modes. The update gate determines the amount of past knowledge that needs to be passed to the future while the reset gate determines the amount of past knowledge to forget. The two gates are vectors that control the flow of information. GRUs are efficient at learning and retaining long-term dependencies in sequential data such as text.

4.3.2 Long-Short Term Memory. While Recurrent Neural Networks are effective in learning temporal patterns, they suffer from a vanishing gradient problem where long term dependencies are lost. A solution to the problem was proposed by Hochreiter et al. [10] by using a variation of RNNs called Long-Short Term Memory (LSTM). LSTMs use a memory cell, to keep track of long-term relationships between text. Using a gated architecture (input, output, and forget), LSTMs are able to modulate the exposure of a memory cell by regulating the gates.

The architecture of the GRU-based model with input embeddings created from the CRAFT corpus is shown in Figure 1. Three types of inputs are used - character, word, and part-of-speech (POS) embeddings. The character inputs are transformed into a 100 dimension embedding while the POS inputs and word embeddings are transformed into NPos (number of unique parts-of-speech) and NTags (number of unique words) dimensions. Character embeddings after a 30% dropout layer are passed through a 150 unit GRU model with a 50% recurrent dropout. The output from this GRU model is concatenated with word and part-of-speech embeddings is passed to a 200 unit bi-directional GRU unit. Out predictions are obtained from this model after a 50% dropout layer. The numerical parameters used in these models were determined after performing a grid search of different settings to arrive at the highest performing parameter combination.

In contrast to the above architecture, in the ELMo GRU model uses a Convolution Neural Network (CNN) for characters in the text in place of the first GRU model (Figure 2). The idea was introduced by Xiang et.al. [24], where CNNs were shown to perform better with low frequency and out of vocabulary words/terms than embeddings. The other inputs remain the same in both models. The output from

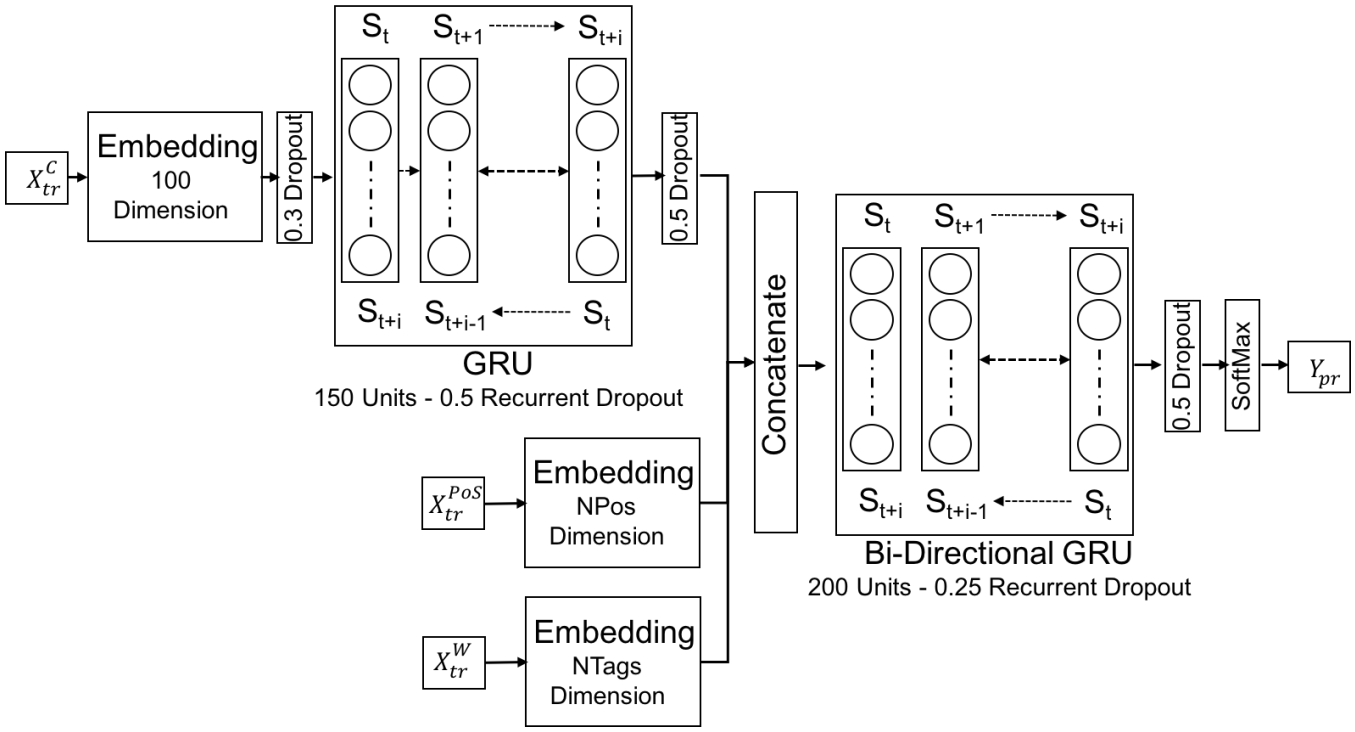

Figure 1: Architecture of a GRU-based model using input character, word, and part-of-speech embeddings created from the CRAFT corpus

the CNN is concatenated with POS embeddings and ELMo word embeddings. This concatenated data is passed to a 200 unit Bi-GRU.

4.4 Encoding formats

Our prior work [13] and along with other studies [8] showed that the input embedding format can impact performance of deep learning models substantially. Here, we combine the above two deep learning models with embeddings generated using the following four techniques to explore the best embedding technique for ontology-powered NER.

GloVe. The Global Vectors for Word Representation (GloVe) is an algorithm that converts words in text to a vector representation [16]. The algorithm trains on word-word co-occurrence statistics from any corpus and results in linear substructures of a word's vector space.

PubMed+PMC. Pre-trained word embeddings from PubMed and PMC with 200 dimensions have been used widely for NER tasks in the biological and biomedical domain [23]. These embeddings are generated using the Word2vec model [14] in the word2vec binary format from a collection of PubMed articles.

ELMo. Embeddings from Language Models (ELMo) [18] is also a vector based encoding system where vectors are derived from a bidirectional LSTM trained on a large corpus. Each token in the encoding contains a representation derived from the complete input sentence resulting in a rich representation with context from the full sentence.

CRAFT. While the previous three embeddings are derived from external corpora of large text, we also developed a domain-specific embedding model trained on the CRAFT corpus. Specifically, instead of utilizing the pre-trained weights as a starting point (for GloVe, PubMed+PMC, and ELMo), the CRAFT embeddings were initialized with random weights and then trained using just the 97 articles in the corpus.

4.5 Incorporating Ontology Hierarchy

The models in this study were designed to take advantage of the hierarchy of concepts in the ontology while predicting annotations. The models are able to make predictions at different levels of specificity in the ontology hierarchy. An ideal/accurate prediction would be the exact annotation in the gold standard. Failing that, the next best prediction would be one of the direct parents of the gold standard annotation followed by the next level parent in the ontology. This intuition follows the true path rule of the Gene Ontology [1] which states that if an object is annotated to a GO term, it is true to assume it is annotated to all subsumers of the GO term. GO terms lose specificity as we travel up the hierarchy towards the root, so it is preferable for the models to predict the most specific subsumer of an annotation when they fail to predict the exact annotation as the gold standard.

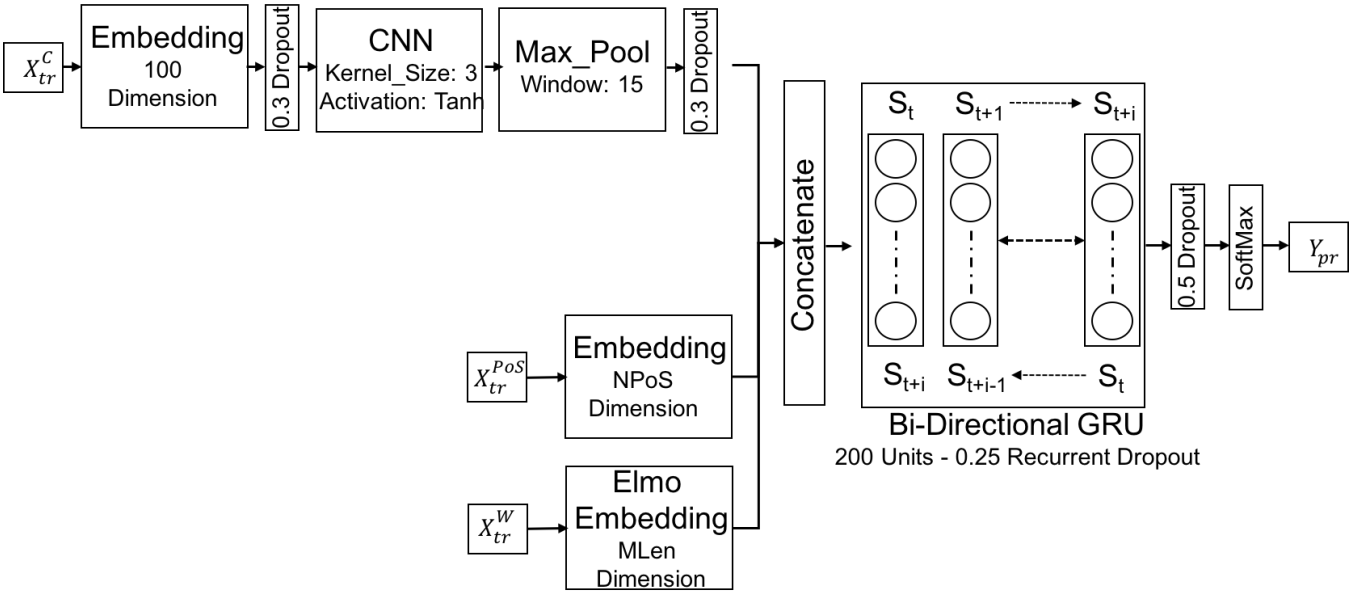

Figure 2: Architecture of a GRU-based model using input character, word, and part-of-speech embeddings created using from a large corpus using Embeddings from Language Models (ELMo)

The ontology hierarchy was incorporated into prediction by creating different levels of inferred annotations from the original CRAFT annotations. Consider the original CRAFT annotations as the "Unmodified (Level 0)" dataset. The Level 1 dataset is generating by replacing every annotation in Level 0 with its immediate parent. In case of concepts with multiple parents, one of the parents is selected. Similarly, data at Level $i+1$ is generated by replacing annotations in level i with their immediate parents. Separate models are trained on each level's data. Here, we created inferred annotation data for four levels. Eight sets of models were created and trained on data from levels 0-4. For each level, a deep learning model was constructed based on the architecture of the top performing model at the lowest level (Table 1).

4.6 Evaluation Metrics

Traditional information retrieval evaluation metrics such as Precision, Recall, and F-1 score were used to evaluate the performance of our modes. In addition, we also use Jaccard similarity - a semantic similarity measure specifically designed to compare ontology concepts [17]. Semantic similarity metrics measure various degrees of relatedness between ontology concepts. In this context, these measures can be used to estimate partial accuracy when the model prediction does not identically match the gold standard annotation. Semantic similarity metrics have been used in various studies to measure inter-annotation agreement and performance of automated annotation text mining tools [7, 13]. Jaccard similarity uses distance between two concepts in an ontology to measure similarity. The closer two concepts are, the more similar they're said to be. The measure is defined as the ratio of the intersection between the ancestors of two concepts to the union of the two sets [17].

5 RESULTS

The CRAFT v3 corpus contains 97 articles containing a total of 7,893 sentences and 16,445 unique words/terms after preprocessing. The corpus is encoded with 1,807 unique GO annotations.

5.1 Model and encoding comparison

First, we compared the performance of two deep learning models: GRU and LSTM (Table 1). Overall, we observed that GRUs outperformed LSTMs. This finding holds across all four encoding formats tested and across two accuracy metrics (F-1 and Jaccard similarity). These results are consistent with our previous findings where bidirectional GRUs outperformed all other models tested.

We used the above two models in conjuction with four different input encodings: CRAFT, PubMed+PMC, GloVe, and ELMo. Of the four encodings, CRAFT encodings showed the best performance (F-1 and Jaccard) followed by ELMo.

5.2 Using the top two predictions of a model

Predictions made by the models are accompanied by a probability score that quantifies the confidence of a prediction. Typically, the prediction with the highest probability is considered as the model's prediction. We explored the possibility of considering the top two predictions based on their probabilities to see if that could improve prediction accuracy. We found that the prediction accuracy (as measured by F-1 score) could be improved by 6-11% for GRUs and 7-9% for LSTMs by considering the top two predictions instead of the top one prediction. Similar improvements were noted for Jaccard similarity as well.

Table 1: Performance evaluation of the GRU and LSTM based models combined with four input encodings

Model	Precision	Recall	F-1	Top two F-1	Jaccard	Top two Jaccard
GRU-CRAFT	0.77	0.83	0.79	0.85	0.69	0.805
GRU-PubMed+PMC	0.72	0.67	0.68	0.77	0.64	0.75
GRU-GloVe	0.71	0.68	0.68	0.79	0.64	0.75
GRU-ELMo	0.77	0.80	0.78	0.84	0.78	0.85
LSTM-Normal	0.72	0.81	0.75	0.82	0.69	0.79
LSTM-PubMed+PMC	0.71	0.68	0.69	0.77	0.65	0.75
LSTM-GloVe	0.66	0.66	0.64	0.73	0.62	0.72
LSTM-ELMo	0.75	0.78	0.75	0.82	0.76	0.84

Table 2: Performance of the GRU-CRAFT when ontology hierarchy was incorporated

Dataset	Precision	Recall	F-1 Score	Jaccard Semantic Similarity
Unmodified	0.779	0.80	0.78	0.78
Level 1	0.79	0.81	0.79	0.77
Level 2	0.79	0.83	0.80	0.77
Level 3	0.82	0.82	0.81	0.79
Level 4	0.84	0.81	0.82	0.79

5.3 Incorporating ontology hierarchy

Augmenting the models with subsumption semantics from the ontology hierarchy showed modest but certain gains in both F-1 score and semantic similarity for the GRU-ElMo, the best model (see Table 1) from our comparisons (Table 2). We see an F-1 increase of 2% when the data was generalized two levels up and a 4% increase when the data was generalized four levels up the ontology. Jaccard semantic similarity showed more modest gains at 1% on a 4-level generalization. Further generalization did not result in improvements.

6 DISCUSSION

We presented eight deep learning architectures built from two models - GRU and LSTM in combination with four encoding formats - CRAFT, PubMed+PMC, ELMo, and GloVe. Our results show that GRU outperforms LSTM models both in terms of F-1 and semantic similarity. In a comparison of encoding formats, the domain-specific CRAFT encodings result in the best performance with GRU models in terms of F-1, but GRU-ELMo results in the highest semantic similarity. This indicates that GRU-CRAFT achieves a high proportion of exact matches and relatively low partial matches. GRU-ELMo on the other hand, performs better with respect to semantic similarity, presumably due to the high number of high quality partial matches. These results indicate that the choice of evaluation metric is important since different metrics capture different aspects of an approach's performance. For ontology-based annotation systems, we deem semantic similarity metrics to be a more appropriate metric since they allow for the notion of partial retrieval of an ontology concept as compared to traditional metrics such as F-1.

Most deep learning and machine learning algorithms consider the prediction with the highest probability to be the result. Surprisingly, our results show that reasonable improvements in accuracy can be achieved simply by considering the top two predictions of the model instead of only the topmost one. This result can be consistently observed across all models and encodings with both F-1 and semantic similarity. This result goes to show that minor parametric modifications to deep learning models and their outputs can substantially impact performance.

We expected to see substantial gains in performance metrics with the incorporation of semantics from the ontology hierarchy. Specifically, we expected to see increases in the semantic similarity measurements since the model has an opportunity to predict partial matches where it fails to make an exact prediction. Surprisingly, this expectation did not hold true in the results. We see a rather modest effect of including ontology subsumers. We posit that the encoding of ontology subsumers in better formats could realize the goal of performance improvement. Representing each annotation as a sequence of ontology terms starting with the annotation itself, followed by its immediate parent, and so on would provide a more comprehensive representation of the ontology hierarchy.

REFERENCES

[1] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, and others. 2000. Gene ontology: tool for the unification of biology. *Nature genetics* 25, 1 (2000), 25–29.

[2] Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, and others. 2012. Concept annotation in the CRAFT corpus. *BMC bioinformatics* 13, 1 (2012), 161.

[3] Lucas Beasley and Prashanti Manda. 2018. Comparison of natural language processing tools for automatic gene ontology annotation of scientific literature. *Proceedings of the International Conference on Biomedical Ontology* (2018).

[4] Imane Boudellioua, Maxat Kulmanov, Paul N Schofield, Georgios V Gkoutos, and Robert Hoehndorf. 2019. DeepPVP: phenotype-based prioritization of causative variants using deep learning. *BMC bioinformatics* 20, 1 (2019), 65.

[5] Mercedes Arguello Casteleiro, George Demetriou, Warren Read, Maria Jesus Fernandez Prieto, Nava Maroto, Diego Maseda Fernandez, Goran Nenadic, Julie Klein, John Keane, and Robert Stevens. 2018. Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature. *Journal of biomedical semantics* 9, 1 (2018), 13.

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[7] Wasila Dahdul, Prashanti Manda, Hong Cui, James P Balhoff, T Alexander Dececchi, Nizar Ibrahim, Hilmar Lapp, Todd Vision, and Paula M Mabee. 2018. Annotation of phenotypes using ontologies: a gold standard for the training and evaluation of natural language processing systems. *Database* 2018 (2018).

[8] Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2019. Using Similarity Measures to Select Pretraining Data for NER. *arXiv preprint arXiv:1904.00585* (2019).

[9] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* 33, 14 (2017), i37–i48.
[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[11] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* (2016).
[12] Chen Lyu, Bo Chen, Yafeng Ren, and Donghong Ji. 2017. Long short-term memory RNN for biomedical named entity recognition. *BMC bioinformatics* 18, 1 (2017), 462.
[13] Prashanti Manda, Lucas Beasley, and Somya Mohanty. 2018. Taking a Dive: Experiments in Deep Learning for Automatic Ontology-based Annotation of Scientific Literature. *Proceedings of the International Conference on Biomedical Ontology* (2018).
[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[15] José Antonio Minarro-Giménez, Oscar Marin-Alonso, and Matthias Samwald. 2014. Exploring the application of deep learning techniques on medical text corpora. *Studies in health technology and informatics* 205 (2014), 584–588.
[16] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162
[17] Catia Pesquita, Daniel Faria, Andre O Falcao, Phillip Lord, and Francisco M Couto. 2009. Semantic similarity in biomedical ontologies. *PLoS computational biology* 5, 7 (2009).
[18] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).
[19] Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*. Springer, 157–176.
[20] Dietrich Rebholz-Schuhmann, Senay Kafkas, Jee-Hyub Kim, Chen Li, Antonio Jimeno Yepes, Robert Hoehndorf, Rolf Backofen, and Ian Lewin. 2013. Evaluating gold standard corpora against gene/protein tagging solutions and lexical resources. *Journal of biomedical semantics* 4, 1 (2013), 28.
[21] Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928* (2017).
[22] Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2018. Cross-type Biomedical Named Entity Recognition with Deep Multi-Task Learning. *arXiv preprint arXiv:1801.09851* (2018).
[23] Zenan Zhai, Dat Quoc Nguyen, Saber A Akhondi, Camilo Thorne, Christian Druckenbrodt, Trevor Cohn, Michelle Gregory, and Karin Verspoor. 2019. Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings. *arXiv preprint arXiv:1907.02679* (2019).
[24] Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *arXiv preprint arXiv:1502.01710* (2015).
[25] Qile Zhu, Xiaolin Li, Ana Conesa, and Cécile Pereira. 2018. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics* 34, 9 (2018), 1547–1554.