Retrofitting Word Embeddings with the UMLS Metathesaurus for Clinical Information Extraction

Mohammed Alawad, S.M.Shamimul Hasan, J. Blair Christian, and Georgia Tourassi
Biomedical Sciences, Engineering, and Computing Group
Health Data Science Institute
Oak Ridge National Laboratory
Oak Ridge, TN 37831 USA
Email: {alawadmm,hasans,christianjb,tourassig}@ornl.gov

***Abstract*—Deep learning has surged in popularity and proven to be effective for various artificial intelligence applications including information extraction from cancer pathology reports. Since word representation is a core unit that enables deep learning algorithms to understand words and be able to perform NLP, this representation must include as much information as possible to help these algorithms achieve high classification performance. Therefore, in this work in addition to the distributional information of words in large sized corpora, we use UMLS vocabulary resources to enrich the vector space representation of words with the semantic relations between words. These resources provide many terminologies pertaining to cancer. The refined word embeddings are used with a convolutional neural (CNN) model to extract four data elements from cancer pathology reports; ICD-O-3 tumor topography codes, tumor laterality, behavior, and histological grade. We observed that using UMLS vocabulary resources to enrich word embeddings of CNN models consistently outperformed CNN models without pre-training word embeddings and even with pre-trained word embeddings on a domain specific corpus across all four tasks. The results show marginal improvement on the laterality task, but a significant improvement on the other tasks, especially for the macro-f score. Specifically, the improvements are 3%, 13%, and 15% for tumor site, histological grade, and behavior tasks, respectively. This approach is encouraging to enrich word embeddings with more clinical data resources to be used for information abstraction tasks from clinical pathology reports.**

***Keywords*-Word embeddings; UMLS; convolutional neural networks; natural language processing;**

I. Introduction

Natural language processing (NLP) approaches offer an effective path to develop automated information extraction tools for cancer registries and provide an efficient and near real-time measurement of cancer incidence, morbidity, survival, and mortality. This is due to the large volumes of pathology reports generated on an annual basis and the variability of reports, as they are sourced from hundreds of healthcare providers and pathology laboratories, which makes the manual processing a time consuming challenge. Deep learning has shown unprecedented success in various artificial intelligence applications including NLP [1]. Among various deep learning algorithms, convolutional neural networks (CNNs) have demonstrated superior performance for document-level information extraction and classification utilizing word embeddings [2].

Word embeddings technique has been adopted by many machine learning and natural language processing applications such as document classification [3], and machine translation [4]. It is an efficient method to convert words to vector representations which are used as inputs of deep learning models. It has shown the ability to capture semantic information via observed similarities in word contexts. Also, embeddings can reduce the dimensionality of feature space as compared to the traditional bag-of-words representation. Conventional word embedding methods, such as word2vec and GloVe, have been widely used for various NLP tasks. However, they need large sized test corpora to provide high quality word vectors. Also, they ignore the valuable information of lexical relations that can be gathered from lexical resources or ontologies, which is very beneficial in the biomedical domain.

Large sized corpora are considered an issue for many applications that have limitations and restrictions in data resources, such as healthcare data. Therefore, pre-trained word embeddings have been used to improve the performance of deep learning models. In this approach, embeddings are first trained on large corpora, such as Wikipedia or Google News, using unsupervised learning. Then, the pre-trained embeddings are used as a layer of deep learning models. Pre-trained word embeddings using word2vec and GloVe have been used in [2], [5] and have increased the accuracy of sentence classification task. However, Pre-trained word

embeddings using word2vec and GloVe did not improve the performance of deep learning models used for clinical applications [6]. In which, word2vec skip-gram is used to learn embeddings from two different sources, a general purpose corpus of articles collected by Google News and a domain specific corpus of biomedical publications hosted on PubMed, but the accuracies were decreased.

Training word embeddings only on unlabeled corpora of text using distributional similarity approach may not distinguish semantic similarity from conceptual association [7]. Therefore, involving extra information about structural features, such as semantic lexicons, has been introduced to enrich word embeddings with words semantic relations. This can be done by a combined objective of distributional representation and structural features [8], [9]. Another approach is to train word vectors using the conventional techniques, and subsequently fine-tune the resultant vectors using the structural representation [10], [11].

In this paper, we use the retrofitting algorithm, presented in [11] to improve the word vector representations. We employed three different vocabulary resources from the UMLS and they are Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT US Edition), National Cancer Institute (NCI) Thesaurus, and International Classification of Diseases, Tenth Revision (ICD-10) [12]. We selected the vocabularies mentioned above because they provide many terminologies pertaining to cancer and researchers are frequently using them for cancer-related text classification [13]–[16].

This paper is organized as follows: Section II presents the related works. Section III discusses the cancer pathology reports and the fundamental four targets that are considered in this paper. It also describes the CNN model used in this paper for automatic information extraction and the word embeddings retrofitting technique. In Section IV we discuss performance evaluation and present the experimental results. Finally, we conclude the paper in Section V.

II. Related Works

Vector space representation of words was early presented in [17]. Then, it has been successfully used in many NLP tasks including language modeling, speech recognition, and machine translation [4]. Word vectors can be constructed by the statistical occurrence of a word in a corpora or by internal representation of words from deep learning models.

In [18], two models were proposed to learn vector representations of words. The goal was to reduce computation complexity and train models on very large datasets. The first architecture is called continuous bag-of-words (CBOW). The second one is called continuous skip-gram. Then, an improvement of skip-gram model was presented in [19] by subsampling of frequent words. It improved the training speed and representation accuracy. Their method can also be used for the CBOW architecture. State-of-the-art word representations for NLP systems use the open source word2vec software package[1] presented in [18], [19]. It can learn high quality vector representations of words and capture semantic and syntactic relations between words precisely.

Recently, researchers have focused on tuning word embeddings by involving extra information, such as Knowledge graphs, to enrich word embeddings and find word relations. This should improve the quality of word vectors that are pre-trained on unlabeled corpora. In [9], graph embedding techniques are merged with word2vec/GloVe word embeddings to enrich them with ontological sources of lexical knowledge. Their proposed method leads to consistent improvement across both the word embedding methods. In [20], enriching word2vec word embeddings with semantic relations obtained from knowledge sources also leads to improvement in accuracy. In [7], word embeddings are enriched with antonymy and synonymy constraints. Their proposed method, which is called counter-fitting, was able to make synonym vector representations more similar, and separate antonym vector representations.

Semantic lexicons were also used to capture semantic information from lexical relational resources and improve word vector representations. In [8], lexical resources are used to update the semantically related vectors by adding an objective while the word vectors are being trained. However, in [11], semantic relation adjustments are applied as a post-processing step. In this method, word vectors can be trained using any method, then the retrofitting algorithm is applied to improve the quality of word vectors.

Combining multiple embeddings in one model has been presented as another approach to improve model performance and word representation. In [2], a multi-channel word embedding method was proposed by concatenating two sets of word vectors. One set is static and pre-trained using word2vec, while the other is fine-tuned during model training. In [21], the improved word vectors (IWV) method was proposed and shown to improve the accuracy of deep learning for sentiment analysis. It concatenates the pre-trained word vectors, using word2vec and GloVe, with vectors resulting from part of speech tagging and vectors resulting from semantic lexicons.

III. Materials and Methods

A. Cancer Pathology Reports

Cancer pathology reports are unstructured free-text documents containing highly descriptive and specific observations of cells and tissues. These reports are a primary information source for cancer registries, and the Surveillance, Epidemiology, and End Results (SEER) program which covers 30% of the US population. In this paper, we used de-identified pathology reports of breast and lung

[1] https://code.google.com/archive/p/word2vec/

cancers, provided from five different SEER cancer registries (Connecticut, Hawaii, Kentucky, New Mexico, Seattle), with the proper IRB-approved protocol. Cancer registry experts manually annotated all pathology reports based on standard guidelines and coding instructions used in cancer surveillance. Their annotations served as the gold standard. Since report sections available in each pathology report vary across pathology labs and registries, we aggregated the text content of every section in the pre-processing phase. Below we describe the tasks with the total number of annotated reports available per information extraction task.

ICD-O-3: We used a corpus of 942 de-identified pathology reports. The corpus contained 12 ICD-O-3 topography codes representing 7 breast and 5 lung primary cancer topographies. For 6 ICD-O-3 codes the dataset included at least 50 reports per code but the remaining 6 ICD-O-3 codes were minimally populated with at least 10 but less than 50 pathology reports per code. Figure 1 illustrates the 12 classes and the corresponding number of annotated reports per code included in the database.

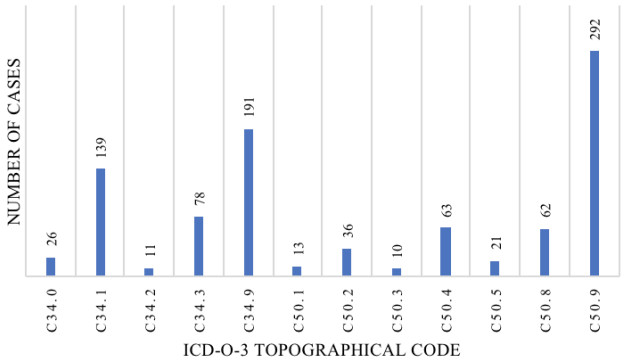

Figure 1: ICD-O-3 topographical codes and number of cases employed in the study.

Tumor Laterality: Laterality in cancer describes which side of a paired organ, breast or lung in this study, is the origin of the primary cancer. Our analysis used a corpus of 815 de-identified pathology reports for which the laterality annotations of 400 reports are clearly stated as right and the laterality of 415 reports is left.

Tumor Behavior: This information describes the behavior of a tumor within a body. We used a corpus of 1033 de-identified pathology reports, that matched to three classes of behaviors. Tumor behaviors and the number of reports available for each class are shown in Figure 2(a). A tumor that does not invade surrounding tissue or spread and grows locally is called benign. If a tumor invades nearby tissue with the potential for spreading, then it is called malignant. While, in situ behavior is a malignant tumor that grows in place.

Cancer Histological Grade: Cancer histological grade is determined by examining the cells and their patterns under a microscope and observing three features: frequency of cell mitosis, tubule formation, and nuclear pleomorphism. It is important to determine how quickly the cells are growing and spreading. In general, a lower grade indicates slower-growing cancer and a higher grade indicates a faster-growing and spreading one. Depending on the extent of the malignancy, a tumor can be graded as 1,2,3, or 4. Our corpus included 645 de-identified pathology reports with ground truth for histologic grade. Cancer histologic grades and the number of reports available for this dataset are shown in Figure 2(b).

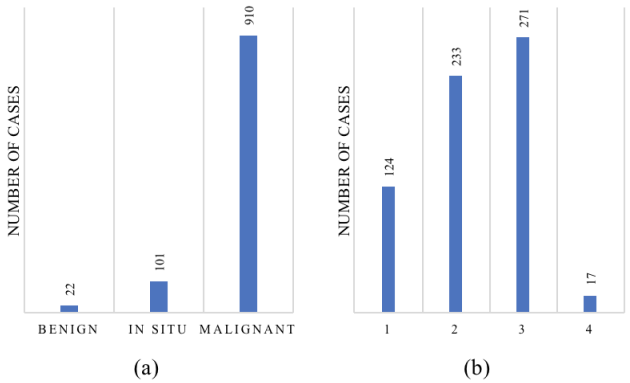

Figure 2: (a) Tumor behavior and number of cases employed in the study. (b) Cancer histologic grades and number of cases employed in the study.

B. UMLS Metathesaurus Vocabularies

The UMLS is a repository of files and software that integrated various biomedical vocabularies and standards developed by the National Library of Medicine (NLM) [22]. The motivation behind the development of the UMLS is to help computer programs to understand the semantics of biomedicine and health [23]. UMLS provides three knowledge sources. First, the Metathesaurus that contains numerous vocabulary terms and codes. Second, the Semantic Network provides semantic types and relations. Third, it contains the Specialist Lexicon for natural language processing [22]. UMLS releases a new version twice in a year (May-AA release and November-AB release) [24]. The latest release of UMLS (2018AA) contains 3,665,926 concepts from 154 sources [25]. As mentioned earlier (in Section I) this paper applied SNOMED CT US Edition, NCI Thesaurus, and ICD-10 vocabularies. We present a brief description of them below.

SNOMED CT US Edition: This provides major general terminology for the electronic health record (EHR). SNOMED CT has hierarchies made from specific conceptual meanings and formal logic-based definitions. Every International Release is made up of concepts, descriptions, and relationships, including work to help implement and

utilize SNOMED CT, subsets, cross maps to previous classifications and coding schemes, with an extensive set of guidelines. The US Edition of SNOMED CT consists of both the International Release of SNOMED CT and the US Extension of SNOMED CT. The US Edition of SNOMED CT provides over three hundred thousand unique concepts. We have more than one million descriptions in SNOMED CT, as well as synonyms to describe a concept. We also have about 903,000 links or semantic relationships between the SNOMED CT concepts. The US Edition of SNOMED CT is a product of the NLM [26].

NCI Thesaurus: This is a product of NCI Enterprise Vocabulary Services (EVS) and in it are the vocabularies about clinical cancer care, translational, and basic research. It consists of public information on cancer together with administrative activity vocabularies. Specifically, it contains definitions, synonyms, and other information on almost ten thousand cancers and related diseases, seventeen thousand single agents and related substances, as well as other topics that associated with cancer and biomedical research. Almost seven hundred new entries are added every month by a team of people from various disciplines [27].

ICD-10: This vocabulary was introduced and is maintained by the World Health Organization (WHO) and is updated every January. It is made up of titles and codes for causes of death, inclusion and exclusion terms for cause-of-death titles, an alphabetical index to diseases and nature of injury, external causes of injury, table of drugs and chemicals, and description, guidelines, and coding rules [28].

C. Word Embeddings

Word embeddings technique has been one of the key breakthroughs in NLP. It is a way of representing a word by a real-valued vector to make algorithms automatically understand analogies of words and capture the semantic properties of words. It is basically a high dimensional feature vector that gives a better representation than one-hot vectors for representing different words. The featurized representations of different words, that will be learned, allow machine learning algorithms to figure out that *word-1* and *word-2* are similar words. Word embeddings can be learned from large text corpora, then transferred to tasks with small datasets. Besides the high quality representation of words, it enables us to use relatively low dimensional feature vectors. Rather than using a *vocab-size* dimensional sparse one hot vector, we can use *word-vector* dimensional dense embedding vector, where *word-vector* $\ll$ *vocab-size*. Implementing an algorithm to learn word embeddings results in an embedding matrix. For example, if our text corpus has V vocabulary words and the feature representation of each word is a d sized *word-vector*, then the embedding matrix is going to be d dimensional by V dimensional. Where each word has a notation corresponding to d by one dimensional embedding vector.

D. Data Preprocessing

1) Pathology Reports: For each pathology report, we extracted the text content and applied standard preprocessing steps to tokenize it. First, we remove XML tags as well as identification tags. In addition, we set all alphabetical characters to lowercase and removed all non-alphanumeric symbols. Then, we created a vocabulary list including all words that have pathology report document frequency of at least two. For each word in the vocabulary list, we assign an embedding vector. Word embedding vectors can be either randomly initialized and fine-tuned via back propagation, initialized from pre-trained word embeddings using a specific corpus, or initialized from retrofitted word embeddings as we will see in Section III-E. Since pathology reports have different lengths, we specify the max pathology report length as the document length L. Reports with a number of tokens less than L will be padded with corresponding tokens [6]. Since each word is represented by a word vector of real numbers $w \in \mathbb{R}^d$, then each document is represented by a document matrix $A \in \mathbb{R}^{d \times L}$

2) UMLS Terms Collection: We downloaded the latest version of the UMLS release from [29] for local installation. We created MySQL database load scripts by using the UMLS installation tool. We generated load scripts with the Active Matathesaurus Subset and followed default setting for all other installation options. We loaded UMLS data into MySQL Community Server 8.0.12. MySQL database is installed on a laptop computer running OS X version 10.12.6 with 2.9 GHz Intel Core i7 processor, 16 GB 2133 MHz LPDDR3 memory and 500 GB of storage space.

We created a python program that reads each word from the vocabulary list (mentioned in Section III-D1) and then connects to MySQL database and fetches the similar UMLS concept names by using the SQL LIKE operator. The python program collected all similar UMLS concept names from three vocabularies mentioned earlier (SNOMED CT, NCI, ICD-10). We experimented with three vocabularies separately and a combination of all of them. Concept names are ordered by the UMLS concept unique identifiers (CUIs). We observe that in the UMLS database sometimes one CUI represents multiple similar concept names. In that situation, we used the first concept name.

E. Retrofitting Word Embeddings

Word embeddings used in this paper are pre-trained in two phases. In phase one, the word embeddings are first trained using traditional distributional similarity approaches without the semantic information from UMLS vocabularies. Then, word vectors corresponding to the set of vocabularies $V = \{v_1, v_2, v_n\}$ from the cancer pathology reports are collected to prepare the pre-trained word embeddings to be refined using semantic relations between words. In the second phase, the retrofitting algorithm, presented in [11], is used to refine the pre-trained word embeddings resulting

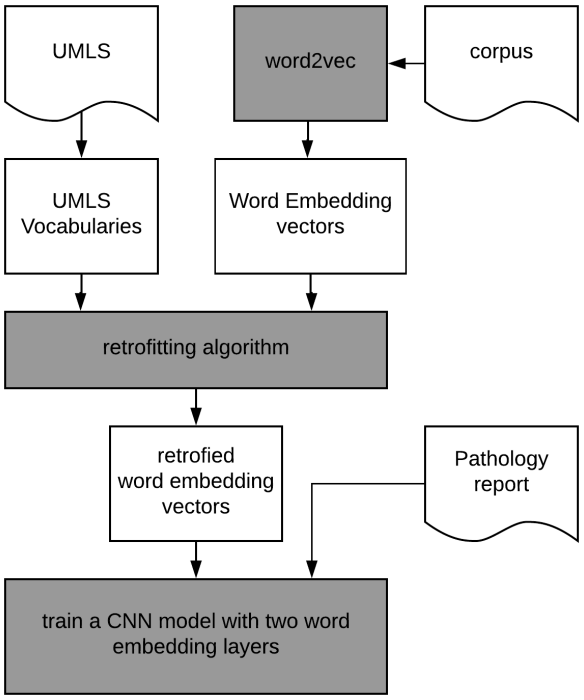

Figure 3: Flow diagram of training a CNN model and retrofitting its word embeddings.

from the first phase. Specifically, a undirected graph is created from the vocabulary words V, the vertices of the graph, and the UMLS semantic relationships between a pair of words, the edges E of the graph. Then, the retrofitted word embeddings $\hat{W}$are initialized to be equal to the word embeddings W resulting from phase one. The objective is to update the word vectors to be close to their counterparts in W and close to the adjacent vertices in the graph. The distance that needs to be minimized to ascertain these objectives can be defined as follows:

$$\Psi = \sum_{i=1}^{n} \left[\alpha_i ||\hat{w} - w||^2 + \sum_{(i,j)\in E} \beta_{ij} ||\hat{w_i} - \hat{w_j}||^2 \right] \quad (1)$$

Where n is the vocabulary list size, w_i is the feature representation of word v_i, $\hat{w_i}$ is the retrofitted feature representation of word v_i, w_j is the feature representation of the adjacent words in the graph to the word v_i, where $(i, j) \in E$. $\hat{W}$ words vectors are updated by taking the first derivative of equation 1 with respect to $\hat{w_i}$ as follows:

$$\hat{w_i} = \frac{\sum_{j:(i,j)\in E} \beta_{ij} \hat{w_j} + \alpha_i w_i}{\sum_{j:(i,j)\in E} \beta_{ij} + \alpha_i} \quad (2)$$

This method can be used to update word embeddings produced from any model. Therefore, we can use any method to pre-train word embeddings, then apply the retrofitting algorithm to update them.

F. Convolutional Neural Networks

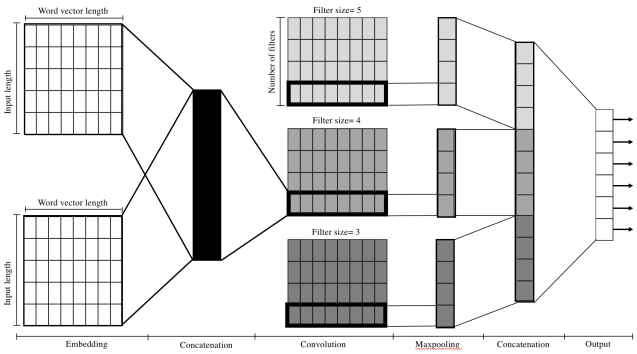

Figure 4: Architecture diagram of a CNN model with two embedding channels used for feature extraction.

CNNs can be applied to the document matrix, as in image processing, by using linear filters with region size l that corresponds to a context length of l word vectors. The convolution layer generates feature maps which are the representation of every context window over the document matrix. The max pooling layer captures the most important features by taking the max value from each feature map as the extracted feature from a particular filter. Then, it aggregates the selected contexts by concatenation. To extract multiple features, we use multiple filters with variable window sizes. The output is connected to a soft-max fully connected layer to produce a rank for each label. For the CNN architecture used in this paper, we applied the hyper parameters presented in [2], [6]. The document length input is defined as 1500 word vectors. The word vector space is 300. The window sizes l of the convolutional filters are 3,4, and 5 with 100 feature maps each. Rectified Linear Unit (ReLU) is used as the activation function. Figure 4 illustrates the architecture diagram of the CNN. The CNN model without pre-training word embeddings has one channel of randomly initialized word embeddings. However, the CNN models, with pre-trained word embeddings or retrofitted word embeddings, have two channels. One channel comes from the pre-trained or retrofitted word embeddings which remain static. While the second channel comes from randomly initialized word embeddings, then fine-tuned during model training.

IV. Performance Evaluation and Experimental Results

In our experiments, we compare our proposed CNN model, that has refined word embeddings using UMLS vocabulary resources, with two baseline architectures, a conventional CNN classifier [6] without pre-training its embeddings, and CNN models with pre-trained embeddings using PubMed corpus.

We implemented a balanced ten-fold cross validation scheme by randomly partitioning the dataset into ten parts with nearly balanced label distributions. For each fold we used one partition once for testing and combined the rest for our training set. We evaluated model performance by aggregating the predicted responses from each test fold. To evaluate the performance accuracy of the three models, we used the standard NLP metrics of micro and macro averaged F-scores. The micro-averaged metrics have class representation roughly proportional to their test set representation, whereas macro-averaged metrics are averaged by class without weighing by class prevalence. Since micro averaged precision, recall, and F-scores are equivalent for multi-class single-label tasks [5], we will use both the micro and macro F- score metric and only the macro average for the recall and precision metrics.

A. Experimental Results

In this work, we test our proposed method on automated information extraction from cancer pathology reports, and specifically extraction of four features, ICD-O-3 tumor site, tumor laterality, tumor behavior, and tumor histological grade. We developed a separate CNN model for each task. They all have the same architecture, but different UMLS vocabulary resources are used to retrofit word embeddings. We used four different UMLS vocabulary resources, SNOMED, NCI, ICD10, and all, which combines all the first three resources. The size of the graph obtained from UMLS vocabulary resources is 3236 nodes and 14675 edges. For all the experiments, we include 95% confidence intervals for each performance metric derived using bootstrapping [30]

In the ICD-O-3 tumor site task, a predictive model must predict the primary subsite of a cancer given the content of the pathology report. The performance results on this task are summarized in Table I. The results show that CNN models with retrofitted word embeddings using any UMLS vocabulary resource consistently outperformed the performance of both CNN models without pre-training word embeddings and with pre-training word embeddings using PubMed corpus. Specifically, the CNN model without pre-training word embeddings micro- and macro-F scores are 0.760 and 0.517, respectively. The CNN model with pre-training word embeddings using PubMed corpus micro- and macro-F scores are 0.763 and 0.519, respectively. Which shows a marginal improvement compared to the CNN model without pre-training word embeddings. The best CNN model with retrofitted word embeddings micro- and macro-F scores are 0.778 and 0.542. This was an improvement of 2% for micro-F score and an improvement of 3% for macro-F score over the baseline models.

Table II shows the performance metrics of the three CNN models on the tumor laterality task. For this task, although retrofitting word embeddings consistently outperformed the two baseline models, the improvement was marginal. This is due to the simplicity of this task and having balanced data representation for both the classes. Where even traditional machine learning techniques can achieve high performance as shown in [31]. Specifically, the CNN model without pre-training word embeddings micro- and macro-F scores are 0.952 and 0.952, respectively. The CNN model with pre-training word embeddings using PubMed corpus micro- and macro-F scores are 0.952 and 0.952, respectively. Which shows no improvement compared to the CNN model without pre-training word embeddings. While the best CNN model with retrofitted word embeddings micro- and macro-F scores are 0.955 and 0.955.

For the tumor behavior task, our proposed method once again outperformed the other two models on all performance metrics. Table III lists the performance comparison results of different classifiers. Specifically, the CNN model without pre-training word embeddings micro- and macro-F scores are 0.921 and 0.649, respectively. The CNN model with pre-training word embeddings using PubMed corpus micro- and macro-F scores are 0.927 and 0.641, respectively. Which shows a slight improvement on micro-F score, but a slight degradation on macro-F score compared to the CNN model without pre-training word embeddings. The best CNN model with retrofitted word embeddings micro- and macro-F scores are 0.947 and 0.749. This was an improvement of 3% for micro-F score and an improvement of 15% for macro-F score over the baseline models.

Finally, our results show that retrofitting word embeddings helps CNN model to predict the histological grade of a cancer given the cancer pathology report better than the models without applying the retrofitting algorithm. Table IV shows the performance comparison metrics of the proposed models and the two baseline models. Specifically, the CNN model without pre-training word embeddings micro- and macro-F scores are 0.758 and 0.684, respectively. The CNN model with pre-training word embeddings using PubMed corpus micro- and macro-F scores are 0.758 and 0.667, respectively. No improvement on micro-F score and a slight degradation on macro-F score compared to the CNN model without pre-training word embeddings were also observed. The best CNN model with retrofitted word embeddings micro- and macro-F scores are 0.811 and 0.776. This was an improvement of 7% for micro-F score and an improvement of 13% for macro-F score over the two baseline models, respectively.

V. Conclusion

In this work, we used different UMLS vocabulary resources to enrich word embeddings with the semantic relations between biomedical terms. The refined word embeddings are used with a CNN model to extract information from cancer pathology reports. We observed that this approach consistently outperformed CNN models with randomly initialized word embeddings or with pre-trained

Table I: Performance comparison of classification models on ICD-O-3 tumor site task.

Model	Micro-F	Macro-F	Precision	Recall
CNN, without pre-train word embeddings	0.760	0.517	0.573	0.497
	0.730, 0.788	0.470, 0.555	0.528, 0.617	0.458, 0.536
CNN, with pre-train word embeddings	0.763	0.519	0.569	0.498
	0.738, 0.791	0.475, 0.557	0.521, 0.619	0.460, 0.539
CNN, retrofitted using SNOMED	0.777	0.540	0.594	0.515
	0.749, 0.803	0.492, 0.580	0.538, 0.640	0.474, 0.556
CNN, retrofitted using NCI	0.775	**0.543**	0.587	**0.522**
	0.747, 0.800	0.497, 0.581	0.540, 0.628	0.479, 0.564
CNN, retrofitted using ICD10	0.774	0.535	0.581	0.514
	0.747, 0.798	0.490, 0.572	0.528, 0.630	0.475, 0.556
CNN, retrofitted using all	**0.778**	0.542	**0.611**	0.516
	0.751, 0.804	0.493, 0.580	0.566, 0.649	0.476, 0.556

Table II: Performance comparison of classification models on tumor laterality task.

Model	Micro-F	Macro-F	Precision	Recall
CNN, without pre-train word embeddings	0.952	0.952	0.952	0.952
	0.936, 0.966	0.936, 0.966	0.936, 0.966	0.936, 0.966
CNN, with pre-train word embeddings	0.952	0.952	0.952	0.952
	0.936, 0.966	0.936, 0.966	0.936, 0.966	0.936, 0.966
CNN, retrofitted using SNOMED	**0.955**	**0.955**	**0.955**	**0.955**
	0.940, 0.968	0.940, 0.968	0.940, 0.968	0.940, 0.968
CNN, retrofitted using NCI	0.953	0.953	0.953	0.953
	0.939, 0.967	0.939, 0.967	0.939, 0.967	0.939, 0.967
CNN, retrofitted using ICD10	0.953	0.953	0.953	0.953
	0.940, 0.967	0.940, 0.967	0.940, 0.967	0.940, 0.967
CNN, retrofitted using all	**0.955**	**0.955**	**0.955**	**0.955**
	0.940, 0.968	0.940, 0.968	0.940, 0.968	0.940, 0.968

Table III: Performance comparison of classification models on tumor behavior task.

Model	Micro-F	Macro-F	Precision	Recall
CNN, without pre-train word embeddings	0.921	0.649	0.744	0.596
	0.904, 0.937	0.569, 0.722	0.634, 0.853	0.531, 0.668
CNN, with pre-train word embeddings	0.927	0.641	0.657	0.633
	0.911, 0.942	0.576, 0.707	0.578, 0.745	0.576, 0.697
CNN, retrofitted using SNOMED	0.945	0.733	0.846	0.684
	0.930, 0.957	0.648, 0.808	0.740, 0.932	0.617, 0.757
CNN, retrofitted using NCI	**0.947**	0.737	**0.860**	0.682
	0.933, 0.960	0.652, 0.813	0.751, 0.945	0.614, 0.756
CNN, retrofitted using ICD10	0.942	0.711	0.809	0.668
	0.927, 0.956	0.621, 0.787	0.688, 0.919	0.601, 0.742
CNN, retrofitted using all	**0.947**	**0.749**	0.836	**0.702**
	0.932, 0.960	0.665, 0.816	0.726, 0.929	0.629, 0.778

Table IV: Performance comparison of classification models on histological grade task.

Model	Micro-F	Macro-F	Precision	Recall
CNN, without pre-train word embeddings	0.758	0.684	0.812	0.640
	0.724, 0.791	0.605, 0.746	0.783, 0.840	0.577, 0.704
CNN, with pre-train word embeddings	0.758	0.667	0.775	0.629
	0.726, 0.791	0.590, 0.736	0.668, 0.837	0.571, 0.696
CNN, retrofitted using SNOMED	**0.811**	0.771	**0.826**	0.741
	0.781, 0.839	0.705, 0.825	0.766, 0.870	0.676, 0.806
CNN, retrofitted using NCI	0.809	0.771	0.828	0.738
	0.778, 0.839	0.700, 0.823	0.762, 0.871	0.669, 0.803
CNN, retrofitted using ICD10	0.805	0.766	0.793	0.746
	0.772, 0.834	0.703, 0.816	0.722, 0.852	0.681, 0.807
CNN, retrofitted using all	**0.811**	**0.776**	0.808	**0.755**
	0.780, 0.842	0.709, 0.828	0.741, 0.860	0.684, 0.820

word embeddings features on the PubMed corpora. The results also demonstrated that combining multiple UMLS vocabulary resources to refine word embeddings has better performance than using one of them. Future direction of this work includes mapping the proposed approach to a bigger dataset, and expanding the scope of our research by integrating various ontologies that include the Human Disease Ontology, Patient Ontology, Experimental Factor Ontology, Symptom Ontology, and Systems Biology Ontology.

ACKNOWLEDGMENT

This work has been supported in part by the Joint Design of Advanced Computing Solutions (JDASC4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health. The authors wish to thank Valentina Petkov of the Surveillance Research Program from the National Cancer Institute and the SEER registries at HI, KY, CT, NM and Seattle for the de-identified pathology reports used in this investigation. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S., Department of Energy under Contract No. DE-AC05- 00OR22725.

REFERENCES

[1] M. R. Vargas, B. S. L. P. de Lima, and A. G. Evsukoff, "Deep learning for stock market prediction from financial news articles," in *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, pp. 60–65, June 2017.

[2] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1746–1751, 2014.

[3] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, (Cambridge, MA, USA), pp. 649–657, MIT Press, 2015.

[4] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Association for Computational Linguistics, 2016.

[5] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 253–263, Asian Federation of Natural Language Processing, 2017.

[6] J. X. Qiu, H. J. Yoon, P. A. Fearn, and G. D. Tourassi, "Deep learning for automated extraction of primary sites from cancer pathology reports," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, pp. 244–251, Jan 2018.

[7] N. Mrksic, D. Ó. Séaghdha, B. Thomson, M. Gasic, L. M. Rojas-Barahona, P. Su, D. Vandyke, T. Wen, and S. J. Young, "Counter-fitting word vectors to linguistic constraints," in *HLT-NAACL*, pp. 142–148, The Association for Computational Linguistics, 2016.

[8] M. Yu and M. Dredze, "Improving lexical embeddings with semantic knowledge.," in *ACL (2)*, pp. 545–550, 2014.

[9] V. Prokhorov, M. T. Pilehvar, D. Kartsaklis, P. Liò, and N. Collier, "Learning rare word representations using semantic bridging," *CoRR*, vol. abs/1707.07554, 2017.

[10] J. Cross, B. Xiang, and B. Zhou, "Good, better, best: Choosing word embedding context," *CoRR*, vol. abs/1511.06312, 2015.

[11] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith, "Retrofitting word vectors to semantic lexicons," in *Proceedings of the 2015 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1606–1615, Association for Computational Linguistics, 2015.

[12] "UMLS Metathesaurus Vocabulary Documentation." URL: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html. [Online; accessed 2018-08-15].

[13] Y. Luo, A. R. Sohani, E. P. Hochberg, and P. Szolovits, "Automatic lymphoma classification with sentence subgraph mining from pathology reports," *Journal of the American Medical Informatics Association*, vol. 21, no. 5, pp. 824–832, 2014.

[14] B. Koopman, G. Zuccon, A. Nguyen, A. Bergheim, and N. Grayson, "Automatic icd-10 classification of cancers from free-text death certificates," *International journal of medical informatics*, vol. 84, no. 11, pp. 956–965, 2015.

[15] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K. Schuler, J. Cooper, W. Guan, and P. C. de Groen, "Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model," *Journal of biomedical informatics*, vol. 42, no. 5, pp. 937–949, 2009.

[16] I. Spasić, J. Livsey, J. A. Keane, and G. Nenadić, "Text mining of cancer-related information: review of current status and future directions," *International journal of medical informatics*, vol. 83, no. 9, pp. 605–623, 2014.

[17] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back propagating errors," vol. 323, pp. 533–536, 10 1986.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[19] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, (USA), pp. 3111–3119, Curran Associates Inc., 2013.

[20] A. Celikyilmaz, D. Hakkani-Tr, P. Pasupat, and R. Sarikaya, "Enriching word embeddings using knowledge graph for semantic tagging in conversational dialog systems," AAAI - Association for the Advancement of Artificial Intelligence, January 2015.

[21] S. M. Rezaeinia, A. Ghodsi, and R. Rahmani, "Improving the accuracy of pre-trained word embeddings for sentiment analysis," *CoRR*, vol. abs/1711.08609, 2017.

[22] "UMLS Quick Start Guide." URL: https://www.nlm.nih.gov/research/umls/quickstart.html. [Online; accessed 2018-08-15].

[23] D. A. Lindberg, B. L. Humphreys, and A. T. McCray, "The unified medical language system," *Methods of information in medicine*, vol. 32, no. 04, pp. 281–291, 1993.

[24] "UMLS Reference Manual." URL: https://www.ncbi.nlm.nih.gov/books/NBK9678/. [Online; accessed 2018-08-15].

[25] "UMLS Statistics - 2018AA Release." URL: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html. [Online; accessed 2018-08-15].

[26] "SNOMEDCT_US - Synopsis." URL: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/SNOMEDCT_US/. [Online; accessed 2018-08-15].

[27] "NCI Thesaurus - Synopsis." URL: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NCI/. [Online; accessed 2018-08-15].

[28] "ICD10 - Synopsis." URL: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/ICD10/. [Online; accessed 2018-08-15].

[29] "UMLS Knowledge Sources: File Downloads." URL: https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html. [Online; accessed 2018-08-15].

[30] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability, Taylor and Francis, 1994.

[31] H.-J. Yoon, A. Ramanathan, and G. Tourassi, *Multi-task Deep Neural Networks for Automated Extraction of Primary Site and Laterality Information from Cancer Pathology Reports*, pp. 195–204. Springer International Publishing, 2017.