I.J. Mathematical Sciences and Computing, **2017, 2, 1-10**
Published Online April 2017 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijmsc.2017.02.01

Available online at ***http://www.mecs-press.net/ijmsc***

Using Deep Learning Towards Biomedical Knowledge Discovery

Nadeem N. Rather [a,*], Chintan O. Patel [a], Sharib A. Khan [a]

[a] *Applied Informatics Inc., New York, NY 10001, USA*

Abstract

A vast amount of knowledge exists within biomedical literature, publications, clinical notes and online content. Identifying hidden, interesting or previously unknown biomedical knowledge from free text resources using an automated approach remains an important challenge. Towards this problem, we investigate the use of deep learning methods that have shown significant promise in identifying hidden patterns from large corpus of text in an unsupervised manner. For example, it can deduce that 'husband' - 'man' + 'woman' = 'wife'. We use the text corpus from MRDEF file in the Unified Medical Language System (UMLS) dataset as training set to discover potential relationships. To evaluate our approach, we cross-verify new relationships against the UMLS MRREL dataset and conduct a manual evaluation from a sample of the non-overlapping set. The algorithm found 32% of new relationships not originally represented in the UMLS. The deep learning methods provide a promising approach in discovering potential new biomedical knowledge from free text.

Index Terms: Biomedical knowledge, Bioinformatics, Deep learning, Machine learning, Unified Medical Language System, UMLS, Word2vec, Word vectors.

1. Introduction

Biomedical knowledge is growing at an unprecedented rate as evident from growing body of research publications, grants, clinical trials and other scientific endeavors. This knowledge in free-form text can be used to identify new knowledge such as potential new drug candidates, molecular targets, new treatments, side-effects and so on. Ever since the first literature based discovery of fish oil's therapeutic association with Raynaud's syndrome [1], several automated approaches [2], [3], [4], [5] have been tried to unlock hidden knowledge from free text. Most often the goal is to identify unknown but potentially interesting relationships that can be then validated by in-vivo medical research or other means. To reduce the cost of the validation efforts, it is important that the algorithms provide results that contain minimum number of false positives. Additionally, if the results can provide insight into the chain of reasoning why a certain fact was marked as

* Corresponding author. Tel.: +91-9906650650
E-mail address: nadeem@trialx.com / nazeer.nadeem@gmail.com

potentially interesting, then it could enable manual evaluation by a medical expert.

To achieve these goals, we approach the literature based discovery problem with deep learning methods. In recent years, the field of deep learning has re-emerged with advanced computational power and fundamental breakthroughs in neural network based learning methods [13]. Deep learning has been applied for various challenging problems including image labelling, speech recognition and robotics. In text analytics, a neural network based algorithm Word2vec was introduced by Mikolov et al [7],[8],[10] that has produced promising results in generating predictions based on a given context of words. It has a couple of interesting properties that make it desirable for our problem:

1. *Unsupervised Approach:* The Word2vec algorithm does not require selection of features or labeling of the dataset. This allows the algorithm to scale to any number of input text documents without requiring any additional effort.
2. *Prediction/Outlier Detection:* Interesting and novel predictions can be generated by performing arithmetic operations on the word vectors to results, for example, the algorithm can deduce that vector('husband') - vector('man') + vector('woman') is close to vector ('wife'). Additionally, the algorithm enables detecting outliers from a group of terms, for example, "cereal" is an outlier in the set "breakfast, cereal, dinner, lunch".

In this paper, we investigate the feasibility of using Word2vec towards literature based knowledge discovery problem. We utilize the Unified Medical Language System (UMLS) resource for training and testing the algorithm and perform a manual evaluation of a sample to investigate its predictive capabilities.

2. Background

In this section we mention brief background related to UMLS [6], [9] and Word2vec [7], [8], [10], [11].

2.1. Word2vec

For Word2vec [7], [8], [10], [11] published by Google in 2013, is a semantic learning framework employing neural network implementation that learns distributed representations for words by deep learning. Word2vec learns swiftly compared to many other modelling tools. Most applications of Word2vec use cosine similarity to quantify closeness. Word2vec is a group of related models that are used to produce so-called word embeddings. These models are shallow, two-layer neural networks that are trained to reconstruct linguistic contexts of words: the network is shown a word, and must guess which words occurred in adjacent positions in an input text. The order of the remaining words is not important (bag-of-words assumption) [8].

After training, Word2vec models can be used to map each word to a vector of typically several hundred elements, which represent that word's relationship to other words. This vector is the neural network's hidden layer [7].

Word2vec can be useful in knowledge discovery applications by identifying terms related to some provided term(s). These relationships captured by word vectors imply that there are some linguistic/semantic regularities between these terms. Some of these relationships may be cross verified from those present in UMLS MRREL dataset, while other relationships are interesting to find, like:

- Which terms are close to meaning.
- How similar are two terms.
- Which term is dissimilar to other terms.
- More interestingly discovering terms by providing terms which are known to be related in some context.

So if we have some known terms: *term1*, *term2*, *term3* and we want to query for unknown *termX* in relative

context of following relationships:

vector(term1) **-** *vector(term2)* **+** *vector(term3)* **=>** ***vector(termX)*** (1)

vector(term1) **is to** *vector(term2)* **as** *vector(term3)* **is to** ***vector(termX)*** (2)

Examples of eq. (1) and (2):

vector('prostate') **-** *vector('male')* **+** *vector('female')* **=>** ***vector('ovarian')***
vector('teeth') is to vector('head') as vector('nose') is to ***vector('neck')***

Thus in above example concepts related to "*prostate*" in males are equivalent to concepts related to "*ovarian*" in females. And likewise if it is known that concept '*teeth'* is related to concept '*head'*, then querying word vector for matching concept for '*nose'* will yield that it is related to '*neck'* in the relative context. To observe strong regularities in the word vector space, it is required to train the models on a large data set, with sufficient vector dimensionality.

2.2. Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS) [6],[9], developed by the US National Library of Medicine (NLM), is perhaps the largest integrated repository of biomedical vocabularies. The 2014AA release of UMLS covers over 2.9 million concepts from more than 150 source vocabularies. Vocabularies integrated in the UMLS Metathesaurus include the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), Consumer Health Vocabulary (CHV), National Center for Biotechnology Information (NCBI) taxonomy, the Medical Subject Headings (MeSH), RxNorm, and International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). UMLS also provides definition dataset MRDEF, there is exactly one row in this file for each definition in the Metathesaurus.

3. Datasets

In this section, we describe the various datasets we employed in training, verifying and testing our model and results.

3.1. Training dataset (UMLS MRDEF)

To train the algorithm, the MRDEF subset of UMLS Metathesaurus was used. A definition is an attribute of an atom (an occurrence of a string in a source vocabulary). A few approach 3,000 characters in length and it contains more than 180,000 concept definitions from around 28 vocabularies.

3.2. Verification dataset (UMLS MRREL)

The MRREL table in the UMLS contains relationships between concepts/terms. There are more than 30 million such relationships in the UMLS MRREL.

3.3. Test dataset

As a test dataset, we generated a set of 30,000 terms with high frequency (see section 5.2) from a text-corpus of data from PubMed abstracts (ncbi.nlm.nih.gov/pubmed), Clinical Trials protocols (clinicaltrials.gov), NIHR grants summary (nih.gov).

4. Approach

In this section we mention algorithms and parameters we varied while building our deep learning models.

4.1. Parameters and algorithms for Word2vec

- *Architecture (learning algorithm)*: Architecture options for Word2vec are skip-gram (slower, better for infrequent words, default) or continuous bag of words (CBOW, fast).
- *Training algorithm:* Hierarchical softmax (better for infrequent words, default) vs negative sampling (better for frequent words, better with low dimensional vectors).
- *Downsampling of frequent words:* Terms that appear with higher frequency in the training data will be randomly down-sampled; default is 0, this parameter can improve both accuracy and speed for large data sets (Google docs recommends values in range of 1e-3 to 1e-5).
- *Word vector dimensionality:* Denotes the number of dimensions (roughly equivalent to topics) present in the vectorial forms, more features result in longer runtimes but not necessarily result in better models.
- *Context / window size:* Denotes terms that occur within a window-neighbourhood of a term, in a sentence, are associated with it during training. Google docs recommend around 10 for skip-gram and around 5 for CBOW.
- *Worker threads:* Number of parallel processes to run, this is machine specific.
- *Minimum word count:* This helps limit the size of the vocabulary to meaningful terms. Terms that occur less than this are ignored in the calculations.

5. Methods & Tools

In this section, we describe the steps that were implemented to generate the vector model, and queries that were performed.

5.1. Word vector generation

We used Word2vec [7],[8],[10],[11] and varied values of its following parameters (explained in Approach section above) to generate word vector:

- *Architecture (learning algorithm)* [LA = skip-gram]
- *Training algorithm [TA = softmax]*
- *Downsampling of frequent words [DFW = le-5]*
- *Word vector dimensionality[WVD = 200/1000/2000]*
- *Context / window size [CWS = 5/10/10]*
- *Minimum word count [MWC = 2]*

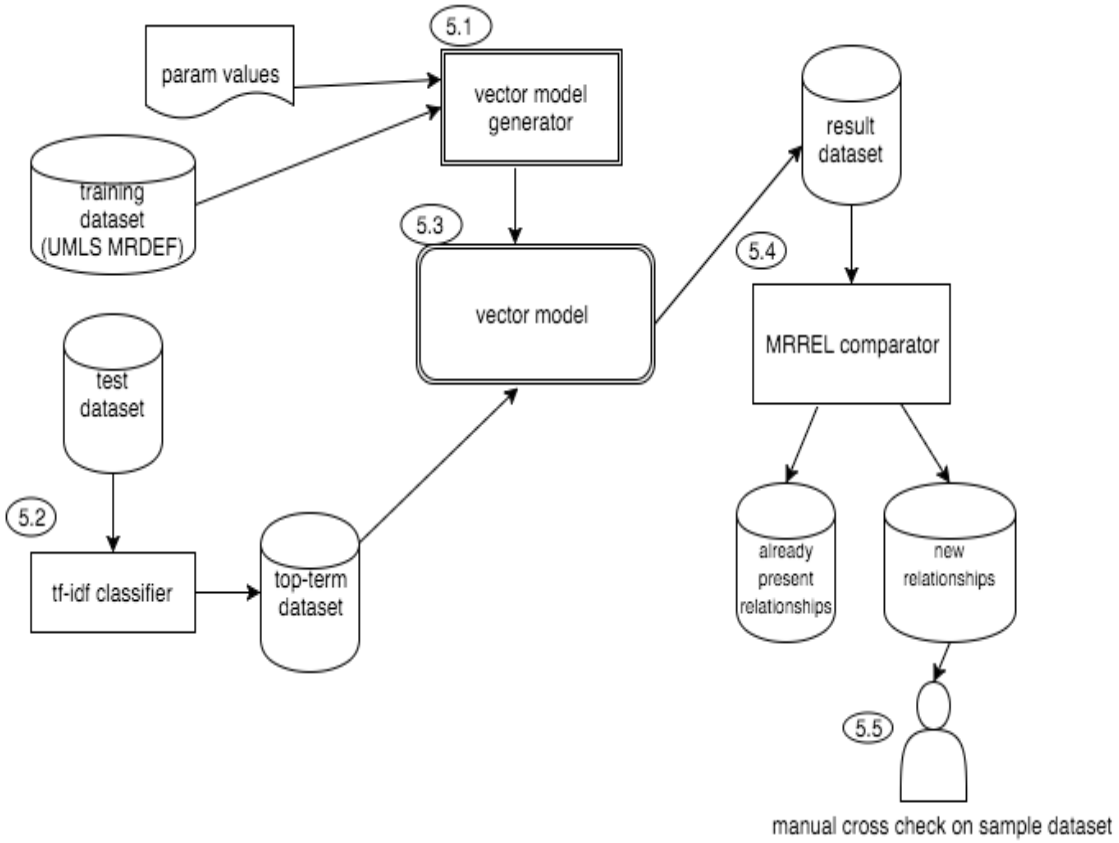

Fig.1. The Experimental Steps Performed to Generate the Model and Verify New Relationships (The Details for Each Step Are Described in Section 5).

5.2. Term frequency–inverse document frequency(tf-idf) on test dataset

We calculated **tf-idf** on our test data set, to find frequent terms which we could use to query our model. We took top 30,000 terms and then took unigrams only as we didn't train the model on phrases and then left out those terms which did not overlap with the terms present in the UMLS.

5.3. Verifying with MRREL

We compared the result set with checking against those relationships present in UMLS MRREL dataset, such that for a given term "x", if it was found in the model then querying the model for its close/similar list of terms. For example, for a set of terms: $[y^1,y^2,y^3,y^4,.....y^n]$, we created a list of pairs from it by cross product of "x" and $[y^1,y^2,y^3,y^4,.....y^n]$:

$$x * [y^1, y^2, y^3, y^4, \dots . y^n] = [(x, y^1), (x, y^2), (x, y^3), (x, y^4), \dots . (x, y^n)] \quad (3)$$

Then

$$P\ (pair\ set) = [(x, y^1), (x, y^2), (x, y^3), (x, y^4), \dots . (x, y^n)] \quad (4)$$

M our matching set will contain those pairs from P (eq. 4) which are related in MRREL:

$$M = \{(x, y): "x" \; is \; related \; to \; "y^i" in \; MRREL\} \quad (5)$$

And the set "N" will contain those pairs from P (eq. 4) which weren't found in MRREL:

$$N = \{(x, y): "x" \; is \; not \; related \; to \; "y^i" in \; MRREL\} \quad (6)$$

5.4. Manual cross checking on sample dataset

One of us (SK), a trained physician, evaluated a sample of 100 random pairs in set N (relations not found in MRREL) to group them into 'Strongly Related', 'Related under some conditions' and 'Seems unrelated'.

5.5. Exploratory Analysis (Vector model querying)

We performed an exploratory analysis of the learnt model by manually looking into the interesting results obtained by performing the following tasks:

- *Analogy Prediction:* Predicting the current word given the context.
- *Outlier Detection:* Identifying outlier from a given set of terms.
- *Vector Arithmetic:* Performing vector additions/subtractions to identify resulting terms.

6. Results

In this section we describe what different types of queries we performed on our models and results we got.

6.1. Overlap Analysis

The results of running the algorithm for overlap analysis are shown in Table 1. We found about 23% overlap with relationships in the MRREL by varying "*Word vector dimensionality*" and "*context/window size*". The results of manual analysis of the set N (relations not found in MRREL) is shown in Table 2 indicating that about 32% strongly related relationships were found that did not existing in the UMLS. Additionally, about 19% of these relationships were related under certain conditions and the rest were deemed to be unrelated.

Table 1. Result Summary of Relations Found and not Found in MRREL after Performing the Overlap Analysis and Varying the Parameters of the Algorithm

Pair set count(P)	Word Vector Dimensionality(WVD)	Context / window size(CWS)	Relations found in MRREL(**M**)(eq. 5)	Relations not found in MRREL(**N**)(eq. 6)
19263	200	5	2993 (15.54%)	16270 (84.46%)
13321	1000	10	3122 (23.44%)	10199 (76.56%)
13474	2000	10	3153 (23.40%)	10321 (76.60%)

Table 2. Manual Evaluation Results on Random Sample Set of 100 Pairs from Set N (Relations Not Found in MRREL)

Relationship strength	Total(%)
Strongly related	32%
Related under some conditions	19%
Seems unrelated	49%

Table 3. Examples of Pairs Evaluated Manually and Corresponding Relationship Strength.

Term1	Term2	Strength
obesity	crohn	*Strongly related*
neoplasms	adenomas	*Strongly related*
cholangiocarcinoma	adenoma	*Strongly related*
myopathy	spastic	*Strongly related*
myocarditis	atopic	*Strongly related*
glaucoma	music	*Related under some conditions*
neutropenia	insomnia	*Related under some conditions*
dengue	chancroid	*Related under some conditions*
rhinitis	diarrhea	*Related under some conditions*
fibromyalgia	dyspnea	*Related under some conditions*
glioblastoma	leiomyosarcoma	*Seems unrelated*
hypoglycemia	aseptic	*Seems unrelated*
rectum	larynx	*Seems unrelated*
amyloidosis	prostate	*Seems unrelated*
larynx	cervix	*Seems unrelated*

6.2. Exploratory Analysis

While performing the prediction task we found many interesting results for the context queries like: *vector (term1)* ***is to*** *vector (term2)* ***as*** *vector (term3) is to vector (termX)****:***

- ***Body Part:*** *'finger' is part of 'hand' as 'toe' is part of* ***'foot'***
- ***Type:*** *'acetaminophen' is as type of 'drug' as 'diabetes' is as type of '****disease'***
- ***Form:*** *'blood' is 'liquid' as 'bone' is* ***'solid'***
- ***Symptom:*** *'fever' is symptom of 'diarrhoea' as 'pain' is symptom of* ***'fatigue'***

6.3. Outlier Detection

Querying for outlier detection we found many interesting results like below:

- *Outlier from [pain, leg, head] is* ***pain.***
- *Outlier from [water solid, blood] is* ***solid.***

- *Outlier from [diabetes, cancer, human] is **human.***
- *Outlier from [nail, hair, head] is **head.***

6.4. Relationship querying on vectors

Querying model with algebraic operations between vectors e.g.: *vector (term1) - vector (term2) + vector (term3) => vector (termX)*, we got results like:

- *Vector('prostate') - vector('male') + vector('female') => **vector('ovarian')***
- *Vector('hair') - vector('head') + vector('leg') => **vector('nail')***
- *Vector('body') - vector('salt') + vector('dehydration') => **vector('walking')***
- *Vector('body') - vector('salt') + vector('exercise') => **vector('temperature')***

7. Discussion

The deep learning methods provide a strong basis for performing literature based knowledge discovery. In our experiments, the overlap with MRREL(23%) was low, however, the manual analysis of the non-overlapping set found several "strongly related" relationships 32% and 19% "related under condition" results, with these two adding up to be more than the unrelated relationships unearthed . The results are promising as it indicates that the algorithm was able to identify "known" knowledge without any prior knowledge or supervision. Within the scope of the experiment, the relationships identified outside the MRREL can be considered as an addition to the existing and in some cases "new" knowledge as UMLS does model a significant number of relationships in the biomedical domain but did not contain some of the unearthed findings. Secondly, the exploratory analysis revealed very interesting set of medically relevant patterns. We believe that the methods of analogy prediction, outlier detection and vector operations are powerful in performing knowledge discovery tasks as they provide a chain of reasoning that can be exploited to trace hidden knowledge in free text resources. The word vector space implicitly encodes many linguistic regularities among words and deep learning can capture a lot of syntactic and semantic information [11], [12].

Our future work includes comparing the deep learning methods against existing literature based discovery methods and using a larger text corpus to find more meaningful and strong patterns. We also envision an interactive semi-automated system that can enable use of the exploratory analysis methods to discover hidden patterns. One potentially interesting application of vector arithmetic can be used to combine word vectors from different types of text resources e.g. combining word vectors from NIH grants summary text corpus with PubMed research publication word vectors to identify potentially novel associations.

8. Conclusion

The deep learning approach using Word2vec showed promising initial results towards performing literature based knowledge discovery. The exploratory analysis methods provide a powerful set of new tools to extract hidden knowledge from text corpuses. The unsupervised nature of algorithm makes it suitable to be executed on large datasets. We evaluated the algorithm in a limited experimental setting, further work is needed on evaluating the algorithm on a large-scale literature-based knowledge discovery task.

References

[1] Swanson DR., Fish oil, Raynaud's syndrome, and undiscovered public knowledge.
[2] Cohen AM, Hersh WR. *A survey of current work in biomedical text mining. Briefings in Bioinformatics.*

2005; 6:57–71.
[3] *Srinivasan P, Libbus B. Mining MEDLINE for implicit links between dietary substances and diseases. Bioinformatics*. 2004; 20:i290–i296.
[4] Smalheiser NR, Swanson DR. *Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicin*e 1998; 57: 149-153.
[5] Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. *Improving literature based discovery support by genetic knowledge integration. Studies in Health Technology and Informatics.* 2003; 95:68–73.
[6] UMLS® Reference Manual, http://www.ncbi.nlm.nih.gov/books/NBK9685/
[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*. In Proceedings of Workshop at ICLR, 2013.
[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. *Distributed Representations of Words and Phrases and their Compositionalit*y. In Proceedings of NIPS, 2013.
[9] Bodenreider O. *The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids* Res. 2004; 32:D267–D270.
[10] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. *Linguistic Regularities in Continuous Space Word Representation*s. In Proceedings of NAACL HLT, 2013.
[11] https://code.google.com/archive/p/word2vec/
[12] Slides about word vectors from NIPS 2013 Deep Learning Workshop: https://drive.google.com/file/d/0B7XkCwpI5KDYRWRnd1RzWXQ2TWc
[13] Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "*Reducing the dimensionality of data with neural networks." Science* 313.5786 (2006): 504-507.

Authors' Profiles

Nadeem Nazeer Rather: *Software & Semantic Data Engineer*, Applied Informatics Inc. Nadeem is a software and semantic data engineer. His areas of expertise include RESTful APIs, biomedical datasets , healthcare IT, software architecture, population analytics, Unified Medical Language System (UMLS). A gold medallist, from the University of Kashmir for his Bachelors in Computer Science. He holds M.Sc. in Information Technology for his masters from Central University Of Kashmir.

Chintan O. Patel: *CTO & Co-founder*, Applied Informatics Inc. Chintan is as a computer scientist and biomedical informatician. His areas of expertise include healthcare IT, software architecture, semantic technologies, machine learning and natural language processing. Previously, he worked at IBM's T.J. Watson Research, on ontology-reasoning and information extraction that then became the infamous Jeopardy winning AI software "Watson". He holds a Ph.D. in Biomedical Informatics from Columbia University

Sharib Ahmad Khan: *Co-founder, Product Lead,* Sharib has a passion for merging technology and healthcare that changes the way we approach healthcare. Sharib holds a Masters in Biomedical Informatics at Columbia University and an MD from University of Medical Sciences, Delhi University

How to cite this paper: Nadeem N. Rather, Chintan O. Patel, Sharib A. Khan,"Using Deep Learning Towards Biomedical Knowledge Discovery", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.3, No.2, pp.1-10, 2017.DOI: 10.5815/ijmsc.2017.02.01