Learning Relations using Semantic-based Vector Similarity

Kinga Budai
Department of Computer Science
Technical University of Cluj-Napoca
Kinga.Budai@student.utcluj.ro

Ioana Bărbănțan
Department of Computer Science
Technical University of Cluj-Napoca
Ioana.Barbantan@cs.utcluj.ro

Mihaela Dînșoreanu
Department of Computer Science
Technical University of Cluj-Napoca
Mihaela.Dinsoreanu@cs.utcluj.ro

Rodica Potolea
Department of Computer Science
Technical University of Cluj-Napoca
Rodica.Potolea@cs.utcluj.ro

***Abstract* — The amount of electronic medical documents is growing rapidly every day. While they carry much information, it becomes more and more difficult to manually process it. Our work represents small steps towards automatic knowledge extraction from medical documents using deep learning and similarity based methods. Our goal here is to identify in an unsupervised manner relations between known medical concepts employing a deep learning strategy with Word2Vec. The current solution requires concepts annotations, as it evaluates the similarities between concepts to identify the relationship between them. The experiments suggest that the strategy we considered (to include the POS as part of the information associated to concepts and relation) represents an important step towards a fully unsupervised learning strategy. Although the POS tags alone are not good enough predictors, the addition of other meta-information and sufficient (quantitative and qualitative) training data may enhance the relation identification process, allowing for a meta learning strategy.**

Keywords—relation extraction; deep learning; data correlation

I. Introduction

In a medical center's database there are a large number of reports describing medical cases of patients having different kind of diseases, diagnoses, examinations, and treatments. These documents can be represented as conceptual maps consisting of medical concepts and the relations between them. The automatic extraction of these relations could provide a further identification of previously unknown relations in documents based on similarities.

Existing reports are usually unstructured and contain complex descriptions of the medical cases. One of the greatest challenges in Natural Language Processing applied on medical fields is to exploit this data and derive a form which allows full utilization of the knowledge. Because there are several supervised and unsupervised Machine Learning based approaches to relation extraction it can be crucial to choose the most adequate one. Deep Learning is such an approach. It is based on a set of algorithms for data processing by applying non-linear functions. Various architectures of Deep Learning are known and used in bioinformatics, like deep neural networks, convolutional deep neural networks, or deep belief neural networks.

In this paper we proposed a solution for analyzing relations that exist between medical concepts represented as N-dimensional vectors constructed by a trained neural network model using deep learning. We identify relations based on the analogy with similar vector pairs in an unsupervised manner.

II. Related Work

Semantic relation extraction and identification tasks are well-known challenges in literature of Text Mining and Natural Language Processing and there is no trivial solution to them. Applied to medical domains they get even more complicated because of the jargon-rich nature of the text. In general two approaches are considered: supervised and unsupervised learning. The goal of supervised learning is generating patterns from a training dataset that can be applied on any other data with similar structure. In the biomedical field this requires a large amount of annotated data, which, most of the time, is not available. On the other hand unsupervised learning clusters and classifies data without human interaction as it discovers their features on its own.

Several challenges concerning medical data processing and text mining are presented in [1]. Some of the approaches dealing with these challenges are detailed in the following. For semantic relation analysis numerous approaches are proposed that use word embeddings. In [11] the authors introduce the TransE method for modeling embedding entities and relationships for multi-relational data representation in low-dimensional vector space. Beside its simplicity the presented method based on translations performed well and efficient in case of large datasets while having a reduced set of parameters. However, TransE demonstrated to have flaws in relations with mapping properties of reflexive and one-to-many/many-to-many/many-to-one relationships. In [12] the authors propose a solution TransH that presents a good trade-off between keeping the efficiency of the TransE model but overcome its simplicity. TransH operates in hyperspace having relations as hyperplanes together with the translation operation on it.

978-1-5090-3899-2/16/$31.00 ©2016 IEEE

The authors in [2] propose a supervised learning framework to determine the existence of a treatment relation between two medical concepts extracted from medical records. For training the framework, a manual annotation of 6864 clinical notes consisting of 958 treatment relations in sentences was performed. The feature vector was composed of lexical features and semantic information that was extracted correlated to the concepts, while the syntactic and structural features are proposed as improvements for the future work. The feature vector is enhanced with information obtained from the MEDication Indication (MEDI) resource and the SemRep rule-based system. Compared to SemRep's performance of 72.34 expressed as F1-measure, the authors report an increment of 12.56%.

Understanding and identifying relations between medical concepts has benefits along the enrichment of the content of ontologies as well. Transforming the textual information into formal information is not an easy task but can be accomplished by complex reasoning and with the help of external systems such as SNOMED-CT, MESH, Medline, Wikipedia, and several articles and datasets available online, as shown in [4]. The authors propose the identification of relations between medical concepts from text and reuse this information to map it into definitions, as formal descriptions. A supervised Machine Learning strategy is used to identify the relations. The dataset is composed of lexical and semantic features, while for the selection of the classification algorithm several experiments were performed using Logistic Regression, Support Vector Machines, Multinomial Naive Bayes, and Random Forests. The best results were obtained in a SVM setup.

The authors in [13] propose a solution of automatic extraction of relations that exist in medical documents. REMed implementation is based on a learning approach using a list of features groups into four categories with the following distribution: lexical (3 features), contextual (6 features), grammatical (4 features) and syntactic (4 features). An analysis of each feature group and their influence is presented. The obtained results are reported using F-measure to compare them with related work.

The 2010 i2b2/VA Workshop on Natural language Processing Challenges for Clinical Records proposed three tasks: a concept extraction task from records of patients, an assertion classification task to assign assertion types for medical concepts and a relation classification task focused on assigning relation types that are between concepts. For these tasks i2b2 and VA provided an annotated reference standard corpus. As described in [4] the relation extraction challenge can be summarized as recognizing three types of relations treatment – problem, test – problem, and problem – problem. Furthermore, the relations treatment – problem and test – problem can be decomposed into subtypes as described in Table I.

III. BACKGROUND

Our objective is to identify relations between concepts in medical documents. In this respect we proposed a solution that employs a processing flow to identify relations between word embeddings based on similarities. The discovered relations could be useful when a new medical case needs to be analyzed. By using historical data when a new examination (tests) needs to be performed, the system can provide a list of actions which were performed in past similar medical cases. On the other hand, when treatment needs to be administered, based on previously known treatment-problem relations we can predict its influence on the problem which is being analyzed.

One approach is based on the strategy implemented in Word2Vec [5]. Word2Vec is a two-layer neural network that processes text in order to translate it into numerical inputs that neural networks can process. It needs to be fed a text input corpus for training where it constructs a set of feature vectors that describe the probabilities of words co-occurrence. Based on similar probabilities calculated this way we can group vectors of related words together in the vector space, thus inferring correlated relations between concepts. The calculated similarities realize a projection of lemmas of words in the input document to an N-dimensional space in which vector representations that are near each other represent similar meanings of the concepts. The results of the model are vectors representing similarities among concepts in the vocabulary. Word2Vec training offers two methods of Deep Learning: using a word to predict the context or using context (surrounding words in the input corpus) to predict the meaning of a word. The first one is known as Skip-Gram model and it performs better than the second approach on large datasets. The prediction of the meaning of a word given its context is known as CBOW (continuous bag of words).

TABLE I. THE SET OF RELATIONS DEFINED IN I2B2 RELATION CHALLENGE

Type 1: treatment-problem relations		
Subcategory abbreviation	Subcategory description	Example
TrIP	Treatment improves problem	*[Solu-Medrol]/tr was given for [thraceal edema]/pr*
TrWP	Treatment worsens problem	*Who presented with [acute coronary syndrome]/pr refractory to [medical treatment]/tr and [TNK]/tr*
TrCP	Treatment causes problem	*[Allergies]/pr included [PENCILIN]/tr and [IODINE]/tr*
TrAP	Treatment administered for problem	*[antibiotic therapy]/tr for presumed [right forearm phlebitis]/pr*
TrNAP	Treatment is not administered because of problem	*He was a poor candidate for [anticoagulation]/tr because of his history of [metastatic Melanoma]/pr*
Type 2: test-problem relation		
TeRP	Test reveals problem	*Patient noted to have [acute or chronic Hepatitis]/pr by [chemistries]/te*
TeCP	Test conducted to investigate problem	*[chest x-ray]/te done to rule out [pneumonia]/pr*
Type 3: problem-problem relation		
PIP	Medical problem indicates medical problem	*[Resting regional wal motion abnormalities]/pr include [mild inferior hypokinesis]/pr*

CBOW architectures are faster at training and have better performance with frequent words. Skip-gram architectures are

more suitable for small amount of data and perform well for rare words. A detailed description of parameters and their influences are described by Mikolov in [6] and summarized in Table II.

TABLE II. PARAMETERS OF THE WORD2VEC MODEL

Parameter	Description
window size	Indicates the distance in tokens between two concepts from the vocabulary that may eventually be correlated.
minimum vocabulary frequency	The number of minimum occurrences of a concept in the text to be considered as part of the vocabulary.
layer size	Size of the hidden layer
negative sampling	Number of contexts randomly selected that are not very similar
down sampling rate	Threshold for occurrence of words

On the other hand, Word2Vec models capture the semantic difference between concepts. A well known example illustrating this feature is: man is to woman as king is to queen that becomes translated in Word2Vec as v(king)-v(man)+v(woman)=v(queen).

Using this behavior we could predict the relation between two concepts: knowing that *R* is a relation between concept *A* and concept *B* and given the semantic difference between the two concepts, we can predict that *R* is also a relation between other two concepts *X* and *Y*, finding their semantic difference to be similar to the one between *A* and *B*. The difference is represented by a vector: semDiff = vector(A) – vector(B). This vector captures the probabilities of both concepts and we label it with a relation *R*.

The relations are represented as tuples of *<C1, R, C2, S>* where *C1* and *C2* are concept pairs (annotated in the describing relation files attached to each record). These tuples have a relation *R* between them and the calculated similarity *S* attached to it.

The input corpus is pre-processed before training: all punctuation marks and numbers are eliminated, the text is converted to lower case and lemmatization is applied. The model has only one hidden layer and it uses weight matrices to translate input to this layer and then to transform it into the output vector. The weights represent probabilities that show for a given word a set of other words which are likely to surround it in the projection space. For example: given a word *W* a vector associated to it is a set of word probability pairs *{C, P}* describing that the word *C* is likely to be around the word *W* with probability *P*. These weights are learned in the training phase and are adjusted in each step using gradient descent algorithm to maximize likelihood. Finally the weights are normalized to sum up to 1.

IV. METHODOLOGY

The strategy we proposed for identifying the relations between concepts consists in applying a deep learning algorithm on a model employing a continuous bag of words (CBOW). The strategy consists of the following phases: pre-processing (input corpus preparation for analysis), training (Word2Vec model building) and testing (evaluating the new relations). Thus the aim is to identify new, unknown relations like in the following example:

v(a heparin drip)-v(atrial fibrillation)+v(zithromax) = v(pneumonia)

The relation type between “a heparin drip” and “artial fibrillation” is known from the annotated data as being of type TrAP. We can draw the conclusion that the same relation type is present between “zithromax” and “pneumonia” based on the simmilarilties on space projection of “a heparin drip” with “artial fibrillation” and “zithromax” with “pneumonia” respectively. Further similar analogies are presented in *Experiments and results* section in Table XI.

A. Pre-processing

Before training, a set of pre-processing steps were applied on the input corpus. Two sets of input files representing summaries were used. One contains free text medical records in English. The other one describes the medical concepts that exist in these records. As the concepts are rather seldom individual words, the challenge we needed to address was the following: we have concepts in relations that are n-grams, i.e. they have more than one word representing the semantic value of the concept. To deal with multiple words concepts we adopted the strategy of building one token from the components: we create artificial unigrams from n-grams (components of compound concepts) by concatenating these words into one single token using a separator defined by us. For example the concept “pulmonary embolus” will be represented as “pulmonary-embolus”. Moreover, they are brought to the root form via lemmatization. For each input file record there is a relation file which describes the relation between medical concepts (concept + relation for training). In this phase the relations and their associated types between concept pairs are extracted.

We need to extract the parts of the records that contain the (annotated) relation and context while avoiding overlearning and noise. We used an intermediary data structure for the input created using a focus on line of interests. For this purpose a window is used on the text extracted from the medical records to take *x* number of phrases before and after the focus relation. The appropriate value of *x* is identified from experiments. This means that the input corpus for the model in the training phase (Training corpus) contains all the lines annotated in the relation file and a number of lines above and below them.

B. Training the model

For training the model two approaches were analyzed: the continuous bag of words (CBOW) and the skip-gram approach. As described in [10] CBOW predicts a word given its context, while skip-gram architectures predict surrounding words given the current word. We performed a preliminary vocabulary experiment to decide which architecture will be

used. We have trained the model using the two approaches and we evaluated the vocabulary based on the number of medical concepts that can be found in it. Based on empirical initial evaluations, we have chosen the CBOW strategy over skip-gram one, for the performance-time trade-off.

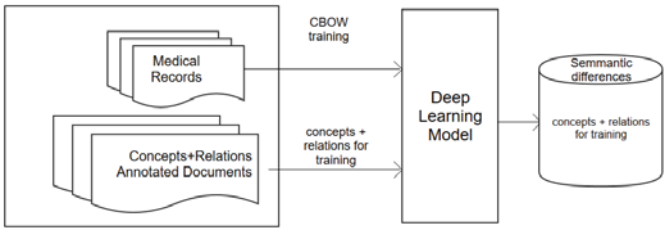

Fig. 1. System architecture for training.

The input parameters were tuned such that the highest accuracy was achieved on the data used for this analysis. At the end of this phase a deep learning model is trained which has the ability to determine the semantic differences associated to relations that are extracted from annotated files that contain concepts and relations.

C. Evaluation

We employed the model using annotated data to identify relations between concepts, on new, unseen documents. The process relies on the semantic difference computed as presented in Fig. 2. For evaluation the same model (Deep Learning model) was used. The Relation Identifier makes associations of relation types and analyzes candidate types for relations. Based on the annotations it evaluates whether the relation assigned is the correct one.

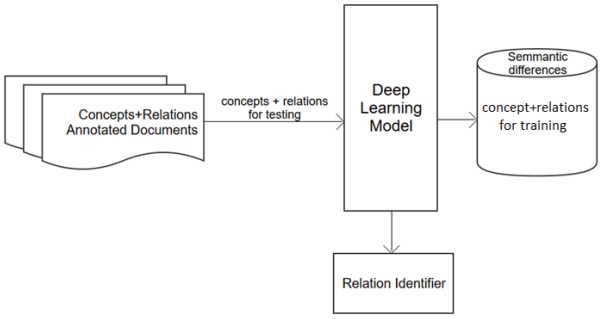

Fig. 2. Evaluation flow

V. Learning the Relations

The core of the system is the interpretation of the semantic difference between concepts in the testing phase. This interpretation is further used to identify new relations in the relation extraction phase, from new, unseen documents. The assignment of the relation types is described in Algorithm 1 where "*R*" is for relation and "*sd*" is for semantic difference. For relation identification we used the similarities between concepts from the vocabulary, computed based on their associated vectors in the N-dimensional search space. If the concept pairs have high cosine similarities (close to 1) it means that their vector representations are very close to each other, thus their semantic value is much alike. Using the trained model we construct the set of semantic differences between relations from the test set (function compute_semantic difference). Then, using these differences we search in the set of annotated relations from the training set for vectors which are most alike to the ones from test set (function get_similar_rels). We store a number of such relations in a list from which candidate relation types will be chosen (function get_relation). It is often the case that candidate types are of different kind so identifying the relation type is not a trivial task. Each relation is given a weight which is adjusted according to the number of occurrences. For example, given such a list having *X* candidates of *type 1* and *Y* candidates of *type 2*, and we know that *X>Y,* (i.e. there are more candidate elements of *type 1*) we assign for the first candidate type a larger weight value, meaning that the relation to be identified is of *type 1* (it has a greater probability than *type 2*).

Algorithm 1:

```
testSet[<C1, R, C2, S>] ← extract(relationset)
similarities ←extractSimilarities(trainSet)
For each e in testSet
begin
    sd ← compute_ semantic_diference (e)
    candidates[<C1, R, C2, S>] ← get_similar_ rels (sd, similarities)
    simRel ← get_ relation (candidates)
    e.R ←simRel.R
end
```

VI. Experiments and Results

The dataset used for evaluations consists of documents provided by Beth Israel-Deaconess Medical Center in Boston and by Partners Healthcare in Boston (resource Medical records from Beth and Partners on Figure no.1) [3]. We further refer to this dataset as resource documents. It consists of 256 medical recordscontaining 6292 medical concepts organized into 3 main categories and 8 subcategories as presented in Table III. Some of these relations represented by two concepts are ambiguous, meaning that between two concepts that are related there exists more than one relation type assigned. The list of relations identified as ambiguous is presented in Table IV.

The available dataset from resource documents containing medical concepts and relations was split into 2 folds in the same way as described in [13]: 170 documents used for training and 256 used for the evaluation. This results in a set of 3118 number of relations for training and a dataset with 6292 of relations used for the evaluation.

For a uniform representation of various occurrences of words (representing concepts and relations in the original documents) to align them in the CBOW approach their lemmas were considered (in this attempt the Stanford CoreNLP tool [7] was employed). Moreover, the part of speech (POS) is considered as tag for a better focus as most of

the times concepts are represented by nouns, noun phrases or adjectives, while relations by verbs, verb phrases or noun phrases. As POS tagger we used Stanford Part-Of-Speech Tagger [8].

TABLE III. ANNOTATED CONCEPTS IN VOCABULARY

Type 1: treatment-problem relations		
Subcategory abbreviation	Subcategory description	Nr of rels
TrIP	Treatment improves problem	203
TrWP	Treatment worsens problem	133
TrCP	Treatment causes problem	526
TrAP	Treatment administered for problem	2616
TrNAP	Treatment is not administered because of problem	174
Type 2: test-problem relation		
TeRP	Test reveals problem	3051
TeCP	Test conducted to investigate problem	504
Type 3: problem-problem relation		
PIP	Medical problem indicates medical problem	2203

For setting up the parameters of the CBOW model, several experiments were conducted, based on parameters description in [9]: we fine tuned the values of the parameters which were fed to the model and we evaluated the results. Their impact is described in Table V.

We evaluated the content of vocabulary based on the number of medical concepts it contains using a set of 150 relations represented by concept pairs instances. For example, adjusting the window size we observed that a suitable value is between 8 and 10, since for larger values the number of instances did not grow significantly without introducing irrelevant correlations of unrelated concepts. The experimental results of adjusting the window size are described in Table VI. Finally, the best results were obtained with the following parameters setup:

- window size: 10
- minimum vocabulary frequency: 1
- layer size (for hidden layer): 400
- negative sampling: 25
- down sampling rate:e-4

The objective of the training step is to obtain for each target concept the similarity vector, represented by tuples of the form <associated_concepts, similarity> of concepts which *are close* to the target concept. Those similarity vectors represent the projection of the concept in an N-dimensional concept space (where N represents the closest N concepts). As the objective is to identify similar concepts, this implies a search for similar vectors in this space. The metric used for this is the cosine similarity between the vectors. Table VII shows pairs of concepts and the relation between them (from annotated documents) and the calculated similarity provided by the model.

In Table VII are shown the negative values, as well. The interpretation of the negative values in our N-dimensional space (N being size of vector) is that the two vectors associated to the corresponding concepts have an angle which is greater than 90 degrees. When training the CBOW model the angles between related words are reduced, thus negative values indicate that concepts from vocabulary are very dissimilar.

TABLE IV. AMBIGUOUS RELATIONS

Concept pairs representing relation	**Record identifier and type of relation**	**Record identifier and a different type of relation**
[Zofran]/tr + [nausea]/pr	In record-124 marked as TrAP	In record 622086964 marked as TrWP
[auscultation]/te + [wheeze]/pr	In 284487129 marked as TeCP	In 130959255 marked as TeRP
[auscultation]/te + [rale]/problem	In 284487129 marked as TeCP	In 130959255 marked as TeRP
[haldol]/tr + [agitation]pr	In 262912613 marked as TrNAP	In 498710998 marked as TrAP

TABLE V. WORD2VEC PARAMETER FINE TUNING IMPACTS

Parameter	**Description**	**Impact**
window size	distance in tokens between correlated concepts	window size too large → irrelevant correlation between word
minimum vocabulary frequency	number of occurrences of a word to be considered as part of vocabulary	if too large → focus on frequent words
layer size	size of hidden layer	more dispersed value for vectors as size is increasing
negative sampling	number of randomly selected contexts that are not very similar	increasing the value we obtained more accurate classification rate
downsampling rate	threshold for occurrence of words	stop words are eliminated if frequency is set to higher value

TABLE VI. FINE TUNING RESULTS FOR WINDOW SIZE

window size	5	8	10	20	30
number of concepts found in vocabulary	52	72	74	103	101

TABLE VII. RELATIONS AND SIMILARITIES OF CONCEPTS

Concept 1	**Concept 2**	**Similarity**	**Type**
intravenous-heparin	pulmonary-embolism	.21	TrAP
anzemet	nausea	.30	TrAP
interval-worsen-in-the-left-lung	multilobar-pneumonia	.85	PIP
you-incision	clear-drainage	.34	PIP
intracranial-pathology	diabetes-insipidus	.49	PIP
urinalysis	febrile	.15	TeCP
recent-ct-scan	destruction-of-femoral-neck	.60	TeRP
the-physical-examination	acute-distress	.57	TeRP
cardiac-exam	a-diffuse-systolic-murmur	.56	TeRP
medication	quite-sedate	-.05	TrCP
percocet	back-pain	-.02	TeRP
levofloxacin	hypotensive	-.01	TrCP

We have evaluated the content of the vocabulary on the training set and the testing set too, by means of the number of medical concepts that are found in it. The obtained results are shown in Table VIII. We have obtained similar accuracy in case of the two sets however the lower accuracy in case of the concepts from training set may indicate some errors.

TABLE VIII. EVALUATION OF VOCABULARY CONTENT ON TRAINING AND TESTING SETS

Set	Accuracy	Total # of concepts	# of identified concepts
Training	97.82	6292	6155
Testing	98.01	3118	3056

We searched in the training set for pairs of vectors which are the most similar to the ones from test set and we mark them as candidate types. We assign a candidate type to a relation only in case the difference of similarities of the concept-pairs (from training and from testing set) is below a threshold. Using this procedure from 6292 relations, 5608 (test set) relations were categorized correctly, this means 89.13% classification rate. The distribution of relations in test set is illustrated in Table IX.

TABLE IX. RELATION DISTRIBUTION IN TEST SET

Relation type	**Number of instances**	**Class distribution in test set (%)**
TrAP	1732	27.52
TrNAP	112	1.78
TrCP	342	5.43
TrIP	152	2.41
TrWP	109	1.73
TeCP	338	5.37
TeRP	2059	32.72
PIP	1448	23.01

The precision, recall and F1-measure of obtained results per subcategories can be seen in Table X. One can observe that the precision and recall varies. For example, the recall is smaller in case of TrIP and TrWP than in case of PIP. This can be justified by the unbalanced nature of our data and the known drawback such data suffer.

We have compared our results to the method proposed in [13] that uses supervised learning for identifying relations in medical documents. We have observed that our results obtained after training the model in an unsupervised manner using the F1-measure metric is higher in case of the TrWP, TrCP, TrAP, TeRP, TeCP, PIP categories, and it is smaller in case of TrIP and TrNAP.

To verify the correctness of candidate types the annotated dataset was used. A part of the results can be seen in Table XI. We say that the relation type was predicted correctly if the value from column Training Relation matches the Testing Relation. The last column (res) shows the evaluation of the predicted relation: F-false, T-true.

TABLE X. COMPARISON OF SUPERVISED AND UNSUPERVISED APPROACH RESULTS

Relation Type	**REMed**			**Our solution**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
TrIP	59.15	27.63	37.66	32.35	21.85	26.08
TrWP	50.0	4.58	8.4	85.41	41	55.40
TrCP	64.80	44.15	52.52	62.5	68.18	65.21
TrAP	85.08	74.76	79.59	93.52	94.65	94.08
TrNAP	56.0	25.0	34.56	94.65	26.85	24.89
TeRP	90.54	79.06	84.41	98.97	96.27	97.60
TeCP	63.05	29.28	40.0	73.99	72.42	73.20
PIP	95.38	66.06	75.71	99.92	97.94	98.28

TABLE XI. RESULTS OF RELATION IDENTIFICATION

Training			Testing			
Concept 1	Concept 2	Relation	Concept 1	Concept 2	Relation	Res
a-pelvic-ultrasound	extensive-fluid	TeRP	blood-culture	febrile	TeCP	F
biopsy	abnormal-vasculature	TeRP	a-elevated-left-hemidiaphragm	splint-on-that-side	PIP	F
a-pelvic-ultrasound	extensive-fluid	TeRP	glucose-level	acutely-agitate	TeCP	F
intravenous-heparin	pulmonary-embolism	TrAP	white-strip	you-incision	TrAP	T
a-pelvic-ultrasound	extensive-fluid	TeRP	mrus-of-the-brain	multiple-focus-susceptibility-artifact-within-the-brain	TeRP	T
you-incision	foul-smell-or-colorful-drainage	PIP	basilar-artery-stenosis	basilar-thrombosis	PIP	T

VII. CONCLUSION AND FURTHER WORK

Our work represents a solution for relations identification and extraction from unstructured medical documents. The current solution relies on the existence of annotated documents to learn from. The annotation assumes the concepts and relations are tagged on the training documents; from here, tuples of the form <concept1, relation, concept2> are extracted, and similarity vectors are learnt. They are further used to find paired concepts in the target documents and to identify relations between them. The addition of meta-information (POS at present) boosts the performance of our solution, so, we claim it is a step towards unsupervised learning.

Our further work considers the addition of other meta-information with this objective. Assuming good quality and enough training data being available, at least a semi-supervised approach is feasible. Furthermore, we attempt to propose and automatic learning method for chosing threshold values (for cosine similarity at relation identification) used to process similarity scores.

REFERENCES

[1] Holzinger A, Schantl J, Schroettner M, Seifert C, Verspoor K. Biomedical text mining: state-of-the-art, open problems and future challenges. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. Berlin: Springer; 2014. p. 271– 300

[2] Cosmin A. Bejan, Joshua C. Denny, Learning to Identify Treatment Relations in Clinical Text, AMIA Annu Symp Proc. 2014; 2014: 282–288. Published online 2014 Nov 14.

[3] Alina Petrova, Yue Ma, George Tsatsaronis, Maria Kissa, Felix Distel, Franz Baader and Michael Schroeder, Formalizing biomedical concepts from textual definitions, Journal of Biomedical Semantics 2015, 6:22, DOI: 10.1186/s13326-015-0015-3© Petrova et al.; licensee BioMed Central. 2015, Published: 2 April 2015.

[4] O. Uzuner , BR. South, S. Shen , et al. 2010. I2b2/VA challenge on concepts, assertions, and relations in clinical text J Am Med Inform Assoc. 2011 Sep-Oct; 18(5): 552–556

[5] DeepLearning4J - Word2Vec: Neural word embeddings in Java http://deeplearning4j.org/word2vec. Last accessed on February 20, 2016.

[6] T. Mikolov, I Sutskever, K. Chen, G. Corrado, J. Dean, Distributed Representation of Words and phrases and their Compositionality, Advances in Neural Information Processing Systems 26, NIPS, 2013

[7] Manning, Christopher D., Surdeanu, Mihai, Bauer, John, Finkel, Jenny, Bethard, Steven J., and McClosky, David. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

[8] K. Toutanova and C. D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

[9] Omar Levy and Yoav Goldberg, Dependency-BasedWord Embeddings, Computer Science Department, Bar-Ilan University, Ramat-Gan, Israel, ACL 2014

[10] T. Mikolov, K. Chen, G. Corrado, J. Dean, “Efficient estimation of word representations in vector space”, International Conference on Learning Representations (ICLR), Scottsdale, Arizona, arXiv: 1301.3781, 2013

[11] Bordes, Antoine, et al. "Translating embeddings for modeling multi-relational data." Advances in Neural Information Processing Systems. 2013

[12] Wang, Zhen, et al. "Knowledge Graph Embedding by Translating on Hyperplanes." AAAI. 2014

[13] Porumb, M., Bărbănțan, I., Lemnaru, C., & Potolea, R. (2015). REMed – Automatic Relation Extraction from Medical Documents. 7th International Conference on Information Integration and Web-based Applications & Services (iiWAS). Brussels.