A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text

Yonghui Wu, Ph.D.[1], Jun Xu, Ph.D. [1],Min Jiang, M.S.[1], Yaoyun Zhang, Ph.D. [1], Hua Xu, Ph.D.[1]

[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

Clinical Named Entity Recognition (NER) is a critical task for extracting important patient information from clinical text to support clinical and translational research. This study explored the neural word embeddings derived from a large unlabeled clinical corpus for clinical NER. We systematically compared two neural word embedding algorithms and three different strategies for deriving distributed word representations. Two neural word embeddings were derived from the unlabeled Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II corpus (403,871 notes). The results from both 2010 i2b2 and 2014 Semantic Evaluation (SemEval) data showed that the binarized word embedding features outperformed other strategies for deriving distributed word representations. The binarized embedding features improved the F1-score of the Conditional Random Fields based clinical NER system by 2.3% on i2b2 data and 2.4% on SemEval data. The combined feature from the binarized embeddings and the Brown clusters improved the F1-score of the clinical NER system by 2.9% on i2b2 data and 2.7% on SemEval data. Our study also showed that the distributed word embedding features derived from a large unlabeled corpus can be better than the widely used Brown clusters. Further analysis found that the neural word embeddings captured a wide range of semantic relations, which could be discretized into distributed word representations to benefit the clinical NER system. The low-cost distributed feature representation can be adapted to any other clinical natural language processing research.

Introduction

Clinical Named Entity Recognition (NER) is to identify boundaries and determine semantic classes (e.g., problems, treatments and lab tests) of clinical concept mentions in clinical text. Over the last few years, much attention has been focused on the clinical NER as it's the critical step of unlocking important healthcare information from narrative clinical text. Much of the important patient information is locked in the narrative clinical text, which is not directly accessible for clinical applications that rely on structured data. Clinical NER systems identify clinical entities from narrative patient reports to support clinical and translational research. Various clinical NER modules have been developed in general clinical Natural Language Processing (NLP) systems (e.g., MedLEE,[1] MetaMap[2], KnowledgeMap[3] and cTAKES[4]). Most of the existing clinical NLP packages are rule-based systems that rely on comprehensive medical vocabularies. Recently, the clinical NLP community organized a series of open challenges with focus on identifying clinical entities from narrative clinical notes, including the 2009 i2b2 (the Center of Informatics for Integrating Biology and the Bedside) challenge[5] on medication information extraction, the 2010 i2b2 challenge[6] on recognizing medical problems, treatments, and tests entities, 2013 Share/CLEF challenge[7] on disorder mention recognition and normalization, and the 2014 SemEval challenge[8] on disorder mention recognition and normalization. Researchers developed rule-based systems, machine learning based systems as well as hybrid systems during the challenges. Currently, most of the state-of-the-art clinical NER systems are primarily based on the machine learning models.[9-11]

Supervised machine learning methods approach the NER as a sequence labeling problem, which aims to find the best label sequence (e.g., BIO format labels) for a given input sequence (individual words from clinical text). Researchers have applied various supervised machine learning algorithms, including Conditional Random Fields (CRFs)[12], Maximum Entropy (ME), and Structural Support Vector Machines (SSVMs)[13], to recognize clinical entities. Among the supervised machine learning algorithms, the CRFs is the most popular one for NER tasks as it's intrinsically designed for sequence labeling problem by modeling the relationships between neighbor tokens. A number of top-ranked NER systems are primarily based on the CRFs. The supervised machine learning algorithms work well as researchers manually extract useful features and feature combinations through feature engineering. Orthographic information (e.g., capitalization of letters, prefix and suffix), syntactic information (e.g. POS tags), n-gram information, semantic information (e.g., UMLS concept unique identifier) and disclosure information (sections in the clinical notes) are often used as features in the typical NER systems. The combination of features, such as the

word combined with POS tags, also prove to be useful.[14] Subsequently, it was identified that the performance of the supervised machine learning algorithm could be further improved by "unsupervised features", which are typically derived from unlabeled corpora using unsupervised machine learning methods such as Brown clustering[15]. Conventionally, Brown clusters are converted into symbolic IDs to form the unsupervised feature representations. Brown clusters have been used successfully in a number of top-ranked clinical NER systems, such as the system from Bruijin[11] in i2b2 2010 challenge and the system from Zhang et al. [10] from the SemEval 2014 Challenge. The study from Tang et al.[14] also found that unsupervised features could improve the identification of clinical entities that not covered by the training corpus. However, this one-hot word representation has limitations in that it only captures a single aspect relation of a word using sparse binary vectors.[16] Researchers have explored the distributional semantics models to derive distributional word representations. Jonnalagadda et al. [17] explored the random indexing model and found that the distributional word representations could enhance the performance of clinical concept extraction. Henriksson et al. [18] further combined the distributional word representations with a large unlabeled in-domain corpus to generate additional features for de-identification of health records.

Recently, there has been an increasing interest in training word embeddings from large unlabeled corpora using neural networks.[19-21] Word embeddings are typically represented as a dense real-valued low dimensional matrix $\boldsymbol{M}$ of size $V \times D$, where V is the vocabulary size and D is the predefined embedding dimension. Each row of the matrix is associated with a word in the vocabulary and each column of the matrix represents a latent feature. The distributed word representations can be derived from the word embeddings. Different from the one-hot word representations such as the clustering feature from Brown clusters, the word embeddings have real-valued numbers to describe multi-aspect relations between words. Usually, the word embedding matrix is first initiated with random values and then tuned using neural networks induced by the neural language model. Bengio[22] and Mikolov[21] proposed different neural networks to train the word embeddings, where the probability of a word given by the previous word was estimated using the cross-entropy criterion. In 2010, Collobert[23] et al. proposed a new neural language model to train word embeddings using ranking loss criteria with negative sampling. The experimental results showed that the ranking based word embeddings derived from the entire English Wikipedia corpus greatly helped the NER task in general English domain.

Previous studies[23,24] have shown that the neural word embeddings could represent abundant semantic meanings and capture multi-aspect relations into a real-valued matrix. However, there is no conclusion on how to use the real-valued word embeddings in machine learning based clinical NER systems. In the biomedical literature domain, Tang et al.[14] conducted a study to evaluate the different types of unsupervised word representations in biomedical NER task. They used the popular word2vec package to generate the word embeddings and showed that the word embedding features improved the F1-score of a baseline NER system by 0.49% (from 70.0% to 70.49%). The Brown cluster features improved the F1-score by 1.2% (from 70% to 71.2%), which was superior to the word embedding features. Tang's study directly used the real values from the embedding matrix as features in a CRFs model without any discretization and the corpus size was relatively small (20,000 sentences from BioCreAtIvE II GM corpus and 22,402 sentences from JNLPBA corpus). Recently, research from Wang and Manning[25] showed that conventional supervised machine learning models, such as the CRFs, have a preference for high dimension discrete feature space instead of low dimension real-valued feature space. Later in 2014, Guo et al.[26] proposed two new strategies for deriving distributed feature representations from neural word embeddings trained from the entire English Wikipedia corpus. The experimental results showed that the proposed binarized embedding features (BinEmb – there are three possible values: "positive", "negative" and "neutral" in BinEmb feature. However, we keep using this name to make it consistent with the previous research) and the distributed prototype features (ProtoEmb) were comparable to the Brown clusters.

However, until now there is no report of using neural word embeddings in the clinical domain. Compared with general English text, the clinical texts are much noisy with frequently occurred ungrammatical sentences, misspellings and abbreviations. It is not clear how the supervised machine learning based clinical NER systems could benefit from the neural word embeddings derived from the noisy clinical corpora. It's also not clear which of the neural word embedding algorithms would be better for clinical NER tasks and how to utilize the word embeddings as features in machine learning based NER systems. In this study, we propose to 1) explore the power of a large unlabeled clinical corpus (403,871 notes) using deep neural networks (DNN); 2) compare the entropy-based neural word embedding algorithm and the ranking-based word embedding algorithm on the clinical NER task; 3) compare three different strategies for deriving distributed word representations from word embeddings in clinical NER tasks. To the best of our knowledge, this is the first study of training neural word embeddings from a large unlabeled clinical corpus and comparing different neural word embedding algorithms and strategies for deriving

distributed word representations. The most related study is that of Guo et al.[26] in the general English domain, where they used only the word embedding algorithm implemented in word2vec. Another related study is by Tang et al.[14] for NER in biomedical literature, where the study directly used the real values as feature weights without discretization, and the corpus size was relatively small.

Methods

Data sets

This study used the annotated corpora from the 2010 i2b2 challenge[6] and the 2014 SemEval challenge[8]. The i2b2 2010 corpus is annotated with three types of clinical entities including Problem, Test, and Treatment. All entities are composed of consecutive words. The SemEval corpus has only one type of entity – the disorder mention. However, the SemEval corpus contains disjoint entities – entities that are composed of more than one piece of text region. The following sentence illustrates an example: "*The **aortic root** and **ascending aorta** are moderately **dilated***". There are two disjoint entities: "*aortic root ... dilated*" and "*ascending aorta ... dilated*". Table 1 shows the detailed information for the two labeled clinical corpora. In order to train the neural word embeddings and the Brown clusters, we utilized a 2.2 gigabytes of unlabeled clinical notes from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II corpus[27]. The MIMIC II corpus is composed of 403,871 notes from four different note types. Table 1 shows the detailed information about the MIMIC II corpus. All the clinical notes were preprocessed using the same pipeline to separate sentences and tokens.

Table 1. Descriptive statistics of data sets used in this study.

Data set		Notes	Entities	Entity types	Note types
i2b2 2010	Training	349	27,837	Problem, Treatment	Discharge
	Test	477	45,009	Test	Progress
SemEval 2014	Training	298	11,156	Disorder	Discharge, radiology
	Test	133	7,971		ECG, ECHO
MIMIC II	N/A	403,871	N/A	N/A	Discharge, radiology ECG, ECHO

ECG: electrocardiogram, ECHO: echocardiogram

The Machine learning-based NER framework

To apply machine learning algorithms to the NER task, the annotated corpora are typically converted into "BIO" format, where "B" denotes "the beginning of a concept", "I" denotes "inside of a concept", and "O" denotes "outside of a concept". E.g., a concept for the semantic category of medical problem can be represented using "B-problem" and "I-problem". In this study, we developed a baseline clinical NER system based on the CRFs model. The baseline system covers the most common NER features including bag-of-word, Orthographic information (word patterns, prefixes and suffixes), syntactic information (POS - part of speech tags), n-gram of word and POS tags (unigrams, bigrams, and trigrams), disclosure information (sections and note types) and combination of words and POSa tags. We used the implementation of CRFs in the CRFsuite package (http://www.chokkan.org/software/crfsuite/). The model parameters were optimized using 5-fold cross validation on the training data, and the best parameters were used to predict the test data.

Integrate word embeddings with the NER framework

Neural word embedding algorithms

We explored two popular word embedding algorithms, including the word2vec from Mikolov[21] and the ranking-based neural word embedding algorithm from Collobert[23]. For word2vec, we used the implementation from "https://code.google.com/p/word2vec/" with the default settings (we used the CBOW model, which is faster and a little bit better than the skip-gram). As there is no out of shelf package for the ranking-based neural word embedding algorithm, we implemented the deep neural network according to the paper from Collobert[23] using Java. We used the suggested parameters to train the neural network with a hidden layer size of 300, a fixed learning rate of 0.01, and an embedding dimension of 50. The standard stochastic gradient descent algorithm was used to optimize the neural network according to the ranking loss. The final word embeddings were represented as dense real-valued matrix. Each row in the embedding matrix associated with a word. Figure 1 shows examples of the word embeddings.

```
atrium   : −0.627 −0.473 0.149 0.165 0.002 −0.015 −0.624 −0.555 0.343 −1.160 …
ventricle: 0.194 −1.492 2.407 0.996 0.379 2.384 −1.808 −0.608 −1.294 0.324 …
pt       : −0.451 0.553 0.399 −0.836 −1.275 1.395 −0.846 0.348 −1.601 −0.484 …
```

Figure 1. Examples of neural word embeddings

Neural word embedding as features

This study compared three different strategies of deriving distributed word representations from neural word embeddings. For each of the strategies, we derived corresponding distributed representations from the MIMIC II corpus and tested the effect of the derived features using the 2010 i2b2 data and the 2014 SemEval data.

1) Raw embedding feature (RawEmb)

The raw embedding feature is a straightforward way of using neural word embeddings. In this method, the real values from the embedding matrix were directly used as feature weights without any post processing. This method will generate the same number (equals to the dimension of the word embeddings) of feature representations for each word. Tang et al. used this strategy in their research[14]. Figure 2 shows examples of the raw embedding features.

```
atrium   :  E0:−0.627 E1:−0.473 E2:0.149 E3:0.165 E4:0.002 E5:−0.015 E6:−0.624 E7:−0.555 …
ventricle:  E0:0.194 E1:−1.492 E2:2.407 E3:0.996 E4:0.379 E5:2.384 E6:−1.808 E7:−0.608 …
pt       :  E0:−0.451 E1:0.553 E2:0.399 E3:−0.836 E4:−1.275 E5:1.395 E6:−0.846 E7:0.348 …
```

Figure 2. Examples of raw embedding features

2) Binarized embedding feature (BinEmb)

The binarized embedding feature was proposed by Guo et al.[26] in 2014 for general English domain. The intuition of the binarized embedding feature is to discretize the real-valued matrix and omit the insignificant dimensions. Given a real-valued neural word embedding matrix $M_{V\times D}$, the binarized embedding features can be derived by converting the real-valued embedding matrix to another discrete-valued matrix $M^*_{V\times D}$ with the discrete symbolic values in [+,−, 0]. For the j^{th} dimension (column) of the embedding matrix, we first calculate the positive mean $MEAN(j)^+$ and negative mean $MEAN(j)^-$ according to the following equations:

$$MEAN(j)^+ = \frac{1}{N_j^+}\sum_{i=0}^{V} M_{i,j}\,,\ M_{i,j} > 0 \tag{1}$$

$$MEAN(j)^- = \frac{1}{N_j^-}\sum_{i=0}^{V} M_{i,j}\,,\ M_{i,j} < 0 \tag{2}$$

Where N_j^+ is the total number of rows with j^{th} column $M.j > 0$, and N_j^- is the total number of rows with j^{th} column $M.j < 0$. Then the discrete-valued matrix M^* can be derived by the following projection:

$$M^*_{i,j} = \begin{cases} +, & \text{if } M_{i,j} > MEAN(j)^+ \\ -, & \text{if } M_{i,j} < MEAN(j)^- \\ 0, & otherwise \end{cases} \tag{3}$$

Using the discrete-valued matrix M^*, we can add the symbolic features according to the row vector associated with the word. Figure 3 shows examples of the binarized embedding features.

```
atrium   :  E0:− E1:0 E2:0 E3:0 E4:0 E5:0 E6:0 E7:0 E8:0 E9:− …
ventricle:  E0:0 E1:− E2:+ E3:+ E4:+ E5:+ E6:− E7:− E8:− E9:0 …
pt       :  E0:0 E1:+ E2:0 E3:− E4:− E5:0 E6:0 E7:0 E8:− E9:0 …
```

Figure 3. Examples of binarized embedding features

3) Distributed prototype feature (ProtoEmb)

The distributed prototype feature was also proposed by Guo et al.[26] for the general English domain. Instead of using the embedding dimensions, the prototype feature method selects prototypical words as representative features for each label. The prototypical words will be assigned as features to the samples according to the distributed similarity

in the embedding matrix. Typically, the prototypical feature words are selected using the normalized pointwise mutual information (PMI) between the word and its labels (equation 4 and 5). For each word '*w*' in the training/test, we calculate the cosine similarity between '*w*' and all the selected prototypical words using the associated embedding vectors. If the cosine similarity is above the predefined threshold, the prototypical word will be assigned as a feature. Following the study by Guo[26], we tested different numbers of prototypical words and cutoff thresholds to select the best parameters. Finally, the top 40 prototypical words and a cutoff threshold of 0.5 were used to assign the prototypical words for the distributed prototype feature. Figure 4 shows the prototypical feature words selected for each label from i2b2 corpus and Figure 5 shows examples of assigned prototypical words.

$$nPMI(label, word) = \frac{PMI(label,word)}{-\ln\, p(label,word)} \quad (4)$$

$$PMI(label, word) = ln\frac{p(label,word)}{p(label)p(word)} \quad (5)$$

B-Problem: afebrile acute hypertension vomiting chills nausea nontender chronic mild some moderate …
I-Problem: disease pain breath failure fibrillation mellitus stenosis effusion infarction distress …
B-Treatment: coumadin lisinopril metoprolol protonix aspirin colace heparin tylenol percocet …
I-Treatment: therapy catheter sulfate drip graft scale bypass saline fluids replacement tube support …
B-test: auscultation glucose hgb abs bun wbc rbc mchc rdw mch plt mcv ast creat cl ptt hct creatinine …
I-test: scan count pressure x-ray ct cultures culture saturation rate biopsy exam bilirubin study fraction …
O: . , : was to with for is she he on and no mg by day as had discharge be has were date history patient …

Figure 4. Examples of prototype words selected using normalized PMI

```
focal    : pleural obstructive dependent acute chronic mild moderate metastatic …
warfarin: coumadin lisinopril metoprolol protonix aspirin colace heparin tylenol percocet …
pt       : patient she he and has which but also there that but chf wbc pain inr …
```

Figure 5. Example of prototype features assigned to word

Experiments and Evaluation

We ran two neural network embedding algorithms and Brown clustering algorithm on the unlabeled MIMIC II corpus to derive word embeddings and Brown clusters. For Brown clustering, we used the implementation from "https://github.com/percyliang/Brown-cluster/" and set number of clusters to 1,000. (We tested the different number of clusters from 50 to 2000 during the 2014 SemEval challenge. The cluster number 1,000 achieved the best performance) For each neural word embedding algorithm, we compared three different strategies for deriving word embedding features. Finally, we combined the best word embedding feature with the Brown clusters to examine how the clinical NER system could benefit from the large unlabeled clinical corpus. The official evaluation scripts provided by the i2b2 organizers and SemEval organizers were used to calculate the strict micro-averaged precision, recall, and F1-score. We report the performances of combining different types of unsupervised word representations on the 2010 i2b2 data and the 2014 SemEval data.

Results

Table 2 and Table 3 show the performances of the CRFs based NER system on the 2010 i2b2 data and the 2014 SemEval data respectively, when using different word representation features. The baseline system achieved F1-scores of 0.799 for the i2b2 data and 0.754 for the SemEval data. The baseline performance on SemEval data is lower than the performance on i2b2 data. The Brown cluster features improved the baseline system by 1.7% for the i2b2 data and 1.3% for the SemEval data. The binarized embedding features outperformed other embedding features and the Brown clusters, by improving the F1-score by 2.3% for the i2b2 data and 2.4% for the SemEval data. The combined feature from the binarized embedding features and the Brown clusters improved the F1-scores by 2.9% for i2b2 data and 2.7% for SemEval data, respectively.

Table 2. Results on the 2010 i2b2 data set.

Features	Precision	Recall	F1-score
Baseline features	0.848	0.755	0.799
+RawEmb (ranking)	0.848	0.768	0.806
+BinEmb (ranking)	0.849	0.797	0.822
+ProtoEmb (ranking)	0.849	0.786	0.816
+RawEmb (word2vec)	0.847	0.766	0.804
+BinEmb (word2vec)	0.846	0.790	0.817
+ProtoEmb (word2vec)	**0.852**	0.782	0.815
+BrownCluster	0.847	0.788	0.816
+BrownCluster+BinEmb (ranking)	0.851	**0.806**	**0.828**

Table 3. Results on the 2014 SemEval data set.

Features	Precision	Recall	F1-score
Baseline features	0.782	0.727	0.754
+RawEmb (ranking)	0.775	0.758	0.767
+BinEmb (ranking)	0.781	0.774	0.778
+ProtoEmb (ranking)	0.784	0.748	0.766
+RawEmb (word2vec)	0.778	0.750	0.764
+BinEmb (word2vec)	0.779	0.764	0.771
+ProtoEmb (word2vec)	**0.789**	0.752	0.770
+BrownCluster	0.778	0.756	0.767
+BrownCluster+BinEmb (ranking)	0.783	**0.780**	**0.781**

All the unsupervised word representation features (including Brown clusters and word embeddings) improved the performances of the clinical NER systems. Further analysis found that the performance improvements are mainly from the recall. For example, the combined features from word embeddings and Brown clusters improved the recall by 5.3% for SemEval data and 5.1% for the i2b2 data, respectively. This is consistent with the previous research from the biomedical literature[14]. The performances of using the two word embeddings are comparable between the two challenge data sets. The ranking based word embedding algorithm from Collobert et al.[23] performed slightly better than the word2vec.

Among the three strategies for using the neural word embeddings, the binarized embedding feature method achieved the best F1-score on both of the challenge data sets. Previous research[14] from the biomedical literature showed that when using a moderate sized corpus (about 42,402 sentences), the Brown cluster feature is superior to the raw word embedding feature (1.2% vs 0.49% on JNLPBA corpus, and 2.1% vs 1.53% on BioCreAtIvE II GM corpus). Our study showed that the binarized embedding feature derived from a much larger corpus (403,871 notes) could be better than the Brown cluster feature (2.3% vs %1.7 for the i2b2 data and 2.4% vs 1.3% for the SemEval data). This could be explained on the basis of the study by Wang and Manning[25], where the authors showed that the discrete high-dimension feature space works better in conventional machine learning models. Another possible reason may be that the marginal benefit of capturing the multi-aspect relations from a big unlabeled corpus is higher than the benefit from a moderate sized corpus. The distributed prototype feature was comparable to the binarized embedding feature. However, the distributed prototype feature benefitted the precision more than the binarized embedding feature.

To examine what the neural word embeddings captured in the real-valued matrix, we calculated the nearest neighbors using the embedding. Table 4 shows several nearest neighbor examples from the ranking based neural word embeddings. We can see that the neural word embeddings capture a wide range of semantic relations in both the general English domain (e.g., number, time unit, verb) and the clinical domain (e.g., gender, modifier, disorder, laterality, body location, medication). The embeddings also captured the semantic relations involving the clinical abbreviations (e.g., yr-year, l-left and r-right).

Table 4. Examples of nearest neighbor words from the ranking based word embeddings.

Word	Top ten nearest neighbors
one	two three four several five six another long 0 large
year	week month weeks yr years days yo old months wk
stopped	restarted initiated discontinued held started begun added weaned titrated diuresed
female	male woman man gentleman ga gestation boy m infant
mildly	moderately markedly slightly severely grossly diffusely somewhat widely relatively extremely
enlarged	prominent edematous thickened widened collapsed dilated opacified atrophic calcified imaged
right	left l r bilateral anterior rt proximal posterior lower upper
atrium	ventricle subclavian arm calf forearm jugular thigh orbit flank elbow
warfarin	allopurinol decadron methadone labetalol hydrochlorothiazide captopril spironolactone fluconazole haldol metformin

This research has limitations. The corpora used in this research are well preprocessed. However, the real world clinical notes are much noisy and heterogeneous. The comparison between neural world embeddings was conducted on a single conventional machine learning model (CRFs). The neural network model usually has high complexity, which may not fit well on the small corpus. We didn't consider the hybrid systems and the existing knowledge bases (such as UMLS). We simplified the system to examine the promise of neural word embeddings – automatically learn knowledge from unlabeled clinical corpora. The best models in the i2b2 challenges (with the best F1-score of 0.852) and SemEval challenge (with the best F1-score of 0.813) show that the existing knowledge and more complex feature combinations could further improve the performance. It's interesting to further examine the world embeddings in the deep neural network based classifiers and combine the existing knowledge bases into word embeddings.

Automatic feature learning from deep neural networks explore rich feature spaces, thus saving the clinical NLP researchers from time-consuming feature engineering. This study showed promising results of using deep neural networks to capture distributed word representations from a large unlabeled clinical corpus to improve the performance of clinical NER systems. Compared with labeled corpus, the unlabeled corpora are much easy to collect. The distributed word embedding features can be adapted to any other clinical NLP system. Moreover, this unsupervised knowledge is low cost - without any involvement of domain knowledge.

Conclusion

This paper studied the neural word embeddings for clinical NER. We systematically compared two popular neural word embedding algorithms and three strategies for deriving distributed word representation features from word embeddings. We also compared the distributed word representation feature with another widely used Brown cluster feature. Evaluation using two challenge datasets showed that the binarized embedding features derived from a large unlabeled corpus could remarkably benefit the clinical NER systems. The word embedding features can be easily adapted to any other clinical NLP research.

Acknowledgement

This study was supported by grants from the NLM 2R01LM010681-05, NIGMS 1R01GM103859 and 1R01GM102282. We would like to thank the 2010 i2b2/VA challenge organizers and the 2014 SemEval challenge organizers for the development of the corpora used in this study.

References

1. Friedman C. Towards a comprehensive medical language processing system: methods and issues. Proc AMIA Annu Fall Symp. 1997:595-599.
2. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. May-Jun 2010;17(3):229-236.
3. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A, 3rd. The KnowledgeMap project: development of a concept-based medical school curriculum database. AMIA Annu Symp Proc. 2003:195-199.
4. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. Sep-Oct 2010;17(5):507-513.

5. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc. Sep-Oct 2010;17(5):514-518.
6. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. Sep-Oct 2011;18(5):552-556.
7. Pradhan S, Elhadad N, South BR, et al. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. J Am Med Inform Assoc. Aug 21 2014.
8. Pradhan S, Elhadad N, Chapman W, Manandhar S, Savova G. SemEval-2014 Task 7: Analysis of Clinical Text. SemEval 2014. 2014;199(99):54.
9. Tang B, Cao H, Wu Y, Jiang M, Xu H. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. BMC medical informatics and decision making. 2013;13 Suppl 1:S1.
10. Zhang Y, Wang J, Tang B, et al. UTH_CCB: A Report for SemEval 2014–Task 7 Analysis of Clinical Text. SemEval 2014. 2014:802.
11. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc. Sep-Oct 2011;18(5):557-562.
12. Lafferty J, McCallum A, Pereira FC. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.
13. Tsochantaridis I, Joachims T, Hofmann T, Altun Y. Large margin methods for structured and interdependent output variables. Paper presented at: Journal of Machine Learning Research2005.
14. Tang B, Cao H, Wang X, Chen Q, Xu H. Evaluating word representation features in biomedical named entity recognition tasks. BioMed research international. 2014;2014:240403.
15. Brown PF, Desouza PV, Mercer RL, Pietra VJD, Lai JC. Class-based n-gram models of natural language. Computational linguistics. 1992;18(4):467-479.
16. Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics; 2010:384-394.
17. Jonnalagadda S, Cohen T, Wu S, Gonzalez G. Enhancing clinical concept extraction with distributional semantics. Journal of biomedical informatics. 2012;45(1):129-140.
18. Henriksson A, Dalianis H, Kowalski S. Generating features for named entity recognition by learning prototypes in semantic space: The case of de-identifying health records. IEEE International Conference on Bioinformatics and Biomedicine. 2014:450-457.
19. Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. Paper presented at: Proceedings of the 25th international conference on Machine learning2008.
20. Mnih A, Hinton GE. A scalable hierarchical distributed language model. Paper presented at: Advances in neural information processing systems2009.
21. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013.
22. Bengio Y, Ducharme R, Vincent P, Janvin C. A neural probabilistic language model. The Journal of Machine Learning Research. 2003;3:1137-1155.
23. Collobert R, Weston J, #233, et al. Natural Language Processing (Almost) from Scratch. J. Mach. Learn. Res. 2011;12:2493-2537.
24. Fu R, Guo J, Qin B, Che W, Wang H, Liu T. Learning semantic hierarchies via word embeddings. Paper presented at: Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics: Long Papers2014.
25. Wang M, Manning CD. Effect of Non-linear Deep Architecture in Sequence Labeling. Proceedings of the Sixth International Joint Conference on Natural Language Processing. 2013:1285-1291.
26. Guo J, Che W, Wang H, Liu T. Revisiting embedding features for simple semi-supervised learning. Paper presented at: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)2014.
27. Saeed M, Villarroel M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. Critical care medicine. May 2011;39(5):952-960.