Extraction of Clinical Timeline from Discharge Summaries using Neural Networks

Adyasha Maharana

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2017

Committee:

Meliha Yetisgen

Adam Wilcox

Program Authorized to Offer Degree:

Biomedical and Health Informatics

University of Washington

Abstract

Extraction of Clinical Timeline from Discharge Summaries using Neural Networks

Adyasha Maharana

Chair of the Supervisory Committee:
Meliha Yetisgen, PhD, Associate Professor
Biomedical and Health Informatics

Discharge summaries are a concise representation of the most important bits of information about a patient's time in the hospital. Converting the free-text into a clinical timeline can facilitate accurate assimilation of information by physicians and the structured data can be used to populate knowledge bases, in clinical decision support systems, etc. Conventional methods for temporal evaluation of discharge summaries employ structured inference and extensive feature engineering. However, they also run the risk of overfitting to the training domain and thus, not being efficient in deployment. Novel methods of natural language processing leverage semantics from large corpuses and produce results with minimum feature engineering. This work explores the use of neural network architectures in clinical entity recognition and temporal evaluation. Recurrent neural networks are found to perform at par with conditional random field systems in clinical entity recognition, scoring 94.04% on the i2b2 2012 dataset. Moreover, they perform better for under-represented entity classes like *'Occurrence', 'Evidential'* and *'Clinical Department'* in a skewed dataset. The out-of-domain evaluation of conditional random fields and neural networks has favorable results on a corpus of ER visit, progress, consult and ICU notes from various medical centers. Neural networks are more agreeable to domain adaptation. This work also explores the use of convolutional neural nets for extraction of within-sentence temporal relations. Preliminary results show that convolutional networks might not be well-suited to the task.

Extraction of Clinical Timeline from Discharge Summaries using Neural Networks

1 Introduction

Hospital discharge summaries are a concise free-text representation of patient's clinical history and care plan. Structured data extracted from this free-text can be useful in several downstream tasks such as visualization, clinical decision-support systems, identification of patient phenotype cohorts, population of knowledge bases etc. A structured dataset consisting of clinical events, time expressions and temporal relations can be used to build clinical timeline of the patient. The i2b2 center has released annotated datasets for this problem and conducted a shared task to accelarate research in this area. Extraction of temporal relations is composed of three subtasks: (1) Detection of clinical events (2) Identification of Temporal Expressions and Normalization (3) Temporal Relation Extraction.

Event detection from clinical notes is a well studied problem in biomedical informatics; yet, it is constantly evolving with novel methods of named entity recognition (NER). The i2b2 2012 dataset is annotated with six entity classes which capture most of the events of importance from a patient's time spent in hospital. Rule based and statistical NLP approaches such as Conditional Random Fields have been used at identifying these entities. These approaches require extensive domain knowledge and feature engineering. (Sun et al., 2013) Similarly, temporal relation extraction modules depend on features from several other NLP modules such as dependency parsing, POS tagging etc. This also leads to the accumulation of errors in the final output. More than half of the temporal relations span across sentences; such long-term dependencies can not be captured with syntactic parsing.

In this paper, we explore discretized word embeddings as new features in structured inference and also implement a neural network architecture for clinical entity recognition. We also compare the ease of domain adaptation between structured inference and recurrent neural networks. We propose a preliminary architecture for extraction of temporal relations from candidate events/time-expressions within sentence with minimum feature engineering.

2 Related Work

The best performing system for clinical event detection on the 2010 i2b2 corpus is a semi-supervised HMM (semi-Markov) model which scored 0.9244 (partial match F1-score) in the concept extraction track (Uzuner et al., 2011). Xu et al. (2012) divided the *Treatment* category into *Medication* and *Non-medication* concepts, and trained two separate conditional random field (CRF) classifiers for sentences with and without medication. With additional features, this system scored 0.9166 on event detection track in 2012 i2b2 challenge, taking the top spot. Tang et al. (2013) built a cascaded CRF system which scored 0.9013 on event detection and came a close second. Sun et al. (2013) showed that these systems found it harder to identify *Clinical Department*, *Occurrence* and *Evidential* concepts. With the surge in deep learning, there have been several new approaches to clinical event detection. Wu et al. (2015) used word embeddings as features in a CRF model and noted improvement in recall for the i2b2 2010 corpus. Chalapathy et al. (2016) implemented a bi-directional LSTM-CRF model with generic embeddings and reported no improvement over the top-performing system in 2010 i2b2 challenge. Jagannatha and Yu (2016a) tested a bi-directional LSTM framework initialized with pre-trained biomedical embeddings on an independent dataset and reported improvement

over a CRF baseline. Recent results show that approximate skip-chain CRFs are more effective at capturing long-range dependencies in clinical text than recurrent neural networks (RNN) (Jagannatha and Yu, 2016b).

The task of identifying temporal expressions in the clinical domain is not unlike that in the general domain. Except for medication dosage frequencies, clinical shorthands and terms like 'post-operative', generic rule-based systems like HeidelTime (Strötgen and Gertz, 2010) are proficient at identifying most temporal phrases from clinical free-text. Sohn et al. (2013) adapted the rules of HeidelTime for clinical corpus and achieved F1-score of 0.9003 on i2b2 data. Xu et al. (2013) used CRF, context-free grammar to tag temporal expressions and scored 0.9144 (overlap match). Most other systems for this task have been built on top of HeidelTime, with additional machine-learning modules. (Sun et al., 2013)

In the i2b2 corpus, any two events in the discharge summary can be a candidate for valid temporal relation. Further, every event is also temporally related to its section time (admission or discharge date), and other temporal expressions within the same sentence. This leads to the inevitable partitioning of relation extraction task into several sub-modules. Tang et al. (2013) developed three separate TLink candidate generation module to detect likely candidates for inter-sentence, intra-sentence and event/section-time temporal relations. These modules are based on outputs from dependency trees, co-reference and rules. After candidate generation, features such as event position, POS tags, verb tense, dependency trees, distance between events, conjunction etc. are used in a combination of CRF and SVM to classify the candidate pairs into temporal relations. With this approach, the system's F-measure was 0.69. (Xu et al., 2013) also use Markov Logic Networks (MLN) to infer those relations which are difficult to classify with SVM. Various other ML-based methods such as the MaxEnt model and rule-based methods also perform competitively. However, performance of relation extraction module still remains the major bottle-neck in an end-to-end system. The state-of-art system for clinical timeline extraction has F1-score of 0.6278 as its end-to-end performance. (Sun et al., 2013).

In the generic domain, deep learning methods have found use in various relation extraction tasks. (Kumar, 2017) Zeng et al. (2014) showed that a convolutional neural network (CNN) is efficient at extracting lexical, sentence-level features and classifying them into relations with the help of position embeddings. Position embeddings are randomly initialized distance vectors, with respect to the distance of current word from both events in the candidate pair. These methods have been verified on component-whole, hyponym, hypernym relationships etc. and on similar tasks in clinical text as well. (Sahu et al., 2016) Dligach et al. (2017) showed that encoding of event containers as xml tags within the input sentence replaces and outperforms position embeddings in classification of temporal relations. However, unlike the variety of temporal relations in i2b2 corpus, this method has been validated on the 'contains' relation only. Peng et al. (2017) have used Graph-LSTMs to extract inter-sentence relations from PubMed articles. In the task Clinical TempEval 2017, extraction of clinical timeline was made even more challenging by posing it as a problem in domain adaptation. (Bethard et al., 2017) Participating systems were trained on notes from colon cancer patients and tested on those from brain cancer patients. Tourille et al. built the best performing system and showed that bi-directional recurrent neural nets (RNN) perform well for extraction of narrative containers. Preventing fine-tuning of pre-trained word embeddings during network training, and replacement of randomly selected event entities with 'unknown' tokens, were the innovative training tactics that facilitated domain adaptation in this system.

We experiment with deep learning methods to build a system for extraction of clinical timeline from i2b2 2012 corpus. For entity recognition, we compare the performance of CRF and RNNs and combine their merits into a hybrid system. We examine the performance of both methods on an out-of-domain dataset. We also develop a neural temporal relation extraction module which goes beyond extraction of narrative containers.

3 Methods

3.1 Dataset

The 2012 i2b2 corpus is made of 310 discharge summaries provided by two medical centers, and consisting of 178, 000 tokens annotated with clinical events, temporal expressions and temporal relations. The training and test sets contain 190 and

120 documents respectively. Each discharge summary has sections for clinical history and hospital course. The number of events, temporal expressions, and TLinks, respectively are 16468, 2368, 33781 in the training set and 13594, 1820, 27736 in the test set. Annotation of clinical events includes problems, tests, treatments, clinical departments, occurrences (admission, discharge) and evidences of information (patient *denies*, tests *revealed*). The inter-annotator agreement for event spans is 0.83 for exact match and 0.87 for partial match (Sun et al., 2013). *Clinical Department* and *Evidential* concepts are under-represented in training set with less than 1000 examples each. Roughly one-third of the TLinks are between pair of events or event and time expression, occurring in the same sentence. The rest are inter-sentence TLinks of which majority are relations between event and section time. The annotations consist of eight temporal classes; however, to enforce transitive closure in the annotations, they are collapsed into three classes - 'Overlap', 'Before' and 'After'. In addition, we use the 2010 i2b2 dataset, which consists of 426 discharge summaries from the same medical centers as the 2012 i2b2 corpus. These notes have been annotated with clinical events only, for the classes 'Problem', 'Test' and 'Treatment'.

The dataset of medical transcription reports from MTSamples consists of 102 discharge summaries and 232 others which includes consult notes, progress notes, ICU, ER visit and physical exam notes. This corpus contains 15089, 4814 and 5518 events tagged as problem, test and treatment respectively. Unlike the i2b2 dataset, these notes do not contain explicit information about the time of admission or discharge. The notes available from MT Samples are crowd-sourced from its users. Hence, it is safe to assume that they are derived from several medical centers, as is evident from the difference in format. We divide this dataset into 210 and 124 notes respectively, for the train and test fold.

3.2 Event Extraction

3.2.1 Baseline

The best performing system in 2012 i2b2 challenge (Xu et al., 2013) requires additional annotation. So, we choose to replicate the second best performing system built by Tang et al. (2013) as our baseline. It is a cascaded CRF classifier, wherein the first CRF is trained on datasets released in 2010 & 2012 to classify for problem, test and treatment. The next CRF is trained on 2012 dataset to extract clinical department, occurrence and evidential concepts. This split in classes is performed to leverage the 2010 dataset which is annotated for the first three classes only. Precision, recall and F-measure (exact event span) for the original system is reported as 93.74%, 86.79% and 90.13% respectively. Our baseline system is built with the same cascaded configuration. The following features are used: N-grams ($\pm$2 context window), word-level orthographic information, syntactic features using MedPOST (Smith et al., 2004), discourse information using a statistical section chunker (Tepper et al., 2012) and semantic features from normalized UMLS concepts (CUIs and semantic types). Tang et al. (2013) employs several other lexical sources and NLP systems for additional features, such as MedLEE, KnowledgeMap and private dictionaries of clinical concepts. For lack of access, they have been left out of our baseline. We have implemented the baseline using CRFSuite package (Okazaki, 2007) and optimum parameters are selected through 5-fold cross-validation on the training set.

3.2.2 Word Embeddings

We use the publicly available source code of GloVe (Pennington et al., 2014) to extract word vectors of dimension 50 for 133,968 words from MIMIC-III. The MIMIC-III dataset (Johnson et al., 2016) contains 2,083,180 clinical notes including discharge summaries, ECG reports, radiology reports etc. Since we are dealing exclusively with discharge summaries in our task, GloVe is run only on the discharge summaries present in MIMIC. These vectors are unfit for direct use in structured prediction and are discretized using methods advocated by Guo et al. (2014).

We also use off-the-shelf biomedical embeddings (dimension=200) which have been made publicly available by Moen and Ananiadou (2013). These embeddings are extracted from abstract and full-text of nearly 22,792,858 PubMed and PMC articles, using the word2vec algorithm. (Mikolov et al., 2013) We experimented with several embedding dimensions in the neural network architecture. To accommodate for values lower than 200, these embeddings are remapped using singular value decomposition (SVD), for only those words which appear in our datasets. Re-

System	TP	FP	FN	Precision	Recall	F1 Score
Baseline	13951	794	2517	**94.63**	84.71	89.40
Baseline + BinEmb	13982	818	2486	94.47	84.90	89.43
Baseline + ProtoEmb	14006	825	2460	94.43	85.06	89.50
Baseline + Brown Clusters	14129	843	2339	94.38	85.78	89.88
Baseline + Brown Clusters + ProtoEmb	14130	860	2338	94.26	85.78	89.82
RNN + random initialization	12370	3123	4098	79.84	75.12	77.38
RNN + MIMIC Embeddings	14315	1373	2153	91.25	**86.93**	89.31
CRF + RNN (Hybrid)	14236	936	2232	93.66	86.45	**89.91**

Table 1: 5-fold cross validation performance of various systems on 2012 i2b2 training set

Entity Class	System	TP	FP	FN	Precision	Recall	F1 Score
Problem	Baseline + Brown Clusters	4607	194	414	**95.96**	**91.72**	**93.79**
	RNN + Embeddings	4429	776	594	85.09	88.17	86.61
Test	Baseline + Brown Clusters	2355	100	242	**95.93**	**90.64**	**93.21**
	RNN + Embeddings	2182	342	415	86.45	83.98	85.20
Treatment	Baseline + Brown Clusters	3469	160	361	**95.62**	**90.57**	**93.03**
	RNN + Embeddings	3296	525	534	86.26	86.06	86.16
Occurrence	Baseline + Brown Clusters	2030	620	1256	76.60	61.78	68.40
	RNN + Embeddings	2042	510	1234	**79.51**	**62.14**	**70.82**
Evidential	Baseline + Brown Clusters	456	116	284	**79.72**	61.62	69.51
	RNN + Embeddings	497	134	243	78.76	**67.16**	**72.5**
Clinical Department	Baseline + Brown Clusters	741	122	256	**85.86**	74.32	79.68
	RNN + Embeddings	813	188	194	79.96	**82.05**	**80.99**

Table 2: Entity-level performance of best performing CRF system and RNN on 2012 i2b2 training set

sults show that these reduced-dimension biomedical embeddings outperform MIMIC-embeddings in clinical entity recognition task. (see Evaluation Metrics & Results)

3.2.3 Recurrent Neural Networks

The bi-directional LSTM-CRF neural architecture introduced by Lample et al. (2016) has been shown to excel on multi-lingual NER tasks. The character embeddings extracted by its bi-directional network, model word prefixes, suffixes and shape - features that are critical to NER. The network is initialized with pre-trained word embeddings which are also fine-tuned during training. Output from the LSTM layer is fed to CRF for joint inference and final decision on sequence tagging.

3.2.4 Hybrid Architecture

The current of state-of-art for detecting problem, test and treatment concepts from clinical text is based on CRF and it has been hard to improve on this baseline, even with neural networks. (Chalapathy et al., 2016) In the first phase of our work, we compare CRF with neural networks and perform error analysis. To this end, two instances of RNN are initialized with MIMIC embeddings; 78.96% words are initialized. First instance is trained to classify problem, test and treatment concepts only and second instance is trained for other three classes. Results from both the networks are merged in a combination module for final evaluation of the end-to-end system. Overlaps are resolved by placing preference on predictions from the first instance. The RNN fails to outperform

Hyperparameter	Value
Hidden layers	1
Word embedding dimension	100
Word hidden layer dimension	100
Character embedding dimension	25
Character hidden layer dimension	50
Dropout	0.5

Table 3: Hyperparameters for RNN architecture

CRF in overall performance, but cross-validation results (presented in Table 2) reveal entity-level differences between CRF and RNN systems. So, we combine the merits of both approaches to create a hybrid end-to-end model. The exact configuration is discussed in the results section.

In later work, we initialize RNN with SVD-reduced PubMed-PMC embeddings and note 3% improvement in F1-score for event detection, at par with the CRF system. Hence, all neural architectures in further experiments are initialized with biomedical embeddings.

3.2.5 Domain Adaptation

With the availability of datasets from two different sources, we are able to examine out-of-domain performance of both, CRF and RNN systems. Since MTSamples notes are annotated for three clinical entity classes only, we focus on evaluation of those three for this task. CRF system with the same feature-set as our baseline is also trained on medical transcriptions from the MTSamples dataset. Two instances of the RNN are initialized with biomedical embeddings and trained on the i2b2 corpus and MTSamples respectively. All systems are tested with in-domain and out-of-domain datasets. For the sake of fair comparison, we present results from i2b2-trained systems on the test-fold of MTSamples dataset. We also created a test fold containing only discharge summaries from MTSamples, but did not find any striking difference between performance on discharge summaries and other notes. Results are presented in Table 6.

3.3 Relation Extraction

3.3.1 Closure of Temporal Relations

The temporal relation classes in i2b2 2012 dataset are 'Before', 'After' and 'Overlap', which can be easily represented in terms of logical operators ¡, ¿ and =. For example, if event A occurs after event B, and event B overlaps with event C, this chain of relations can be expresse as 'A¿B=C'. It is evident from the logical representation, that the dataset needs to have transitive closure of relations in order to be accurate. This means that, the relation 'A¿C' also needs to be a part of the dataset. All inverse relations should also be included in the dataset so that the classifier does not treat them as unrelated event/time expression pairs. From our example, 'B¡A' should be annotated as a temporal relation as well. Before performing relation extraction, we modify the TLink dataset to enforce transitive closure. In addition, all negative instances i.e. temporally unrelated event or time expression pairs, are also explicitly added to the dataset under the class 'Other.

3.3.2 Convolutional Neural Networks

The convolutional network architecture used in this work is inspired from Zeng et al. (2014) and has been modified to suit our task. It is composed of embedding layer, convolutional layer, max-pooling layer, fully-connected layer and softmax layer in that order.

The embedding layer includes position, POS tag, word and event/time expression type embeddings. Except for word embeddings, all others are randomly intialized before training and fine-tuned with the network. There are two position embeddings, one each for representing proximity the two events. Word embeddings are initialized with biomedical embeddings. At the word level, all such embeddings are combined to form one vector. Further, all such embeddings are concatenated at the sentence level before being fed to the convolutional layer.

The convolutional layer is made of filters of various sizes. Features from different context window lengths are captured with varying filter sizes. We experiment with filters of sizes ranging from 2 to 5 and settle with a combination of filter sizes [2, 3, 4]. The output from this layer varies in dimension depending on the sentence length. To counter this variation, the max-pooling layer pools outputs from all word position along every dimension of the feature vector and combines them into a one-dimensional sentence-level vector. This vector is then relayed to a fully-connected layer, with as many outputs as there are classes in the relation classification problem. The final layer in this architecure is the softmax layer, which provides prediction probabilities for the task.

We implement this convolutional network architecture for extracting temporal relations between events and/or time-expressions occurring within the same sentence. It is done in two phases: 1. as a classification task without negative instances 2. as an extraction task with negative instances. The results are presented in Table 7.

4 Evaluation Metrics and Results

We report the micro-averaged precision, recall and F1-score, for 'overlap' match of event spans as

System	TP	FP	FN	Precision	Recall	F1 Score
Tang et al. (2013)	-	-	-	93.74	86.79	90.13
Baseline + Brown Clusters	11664	647	1930	**94.74**	85.80	90.05
Hybrid CRF-RNN	11985	875	1609	93.20	**88.16**	**90.61**

Table 4: Performance of best performing CRF and Hybrid CRF-RNN on 2012 i2b2 test set

Entity Class	System	TP	FP	FN	Precision	Recall	F1 Score
Occurrence	Baseline + Brown Clusters	1509	489	991	**75.53**	60.36	67.10
	Hybrid	1565	563	935	73.54	**62.60**	**67.63**
Evidential	Baseline + Brown Clusters	370	76	226	**82.96**	62.08	71.02
	Hybrid CRF-RNN	446	177	150	71.59	**74.83**	**73.17**
Clinical Department	Baseline + Brown Clusters	557	109	176	**83.63**	75.99	79.63
	Hybrid CRF-RNN	657	234	76	73.74	**89.63**	**80.91**

Table 5: Entity-level performance of best performing CRF and Hybrid CRF-RNN on 2012 i2b2 test set

per the i2b2 evaluation script. TP, FP, FN counts of overall performance are calculated for entity spans, irrespective of entity tag. Systems are also evaluated for performance in individual entity classes and TP, FP, FN counts are compared between the CRF and RNN+Embedding systems. We perform five-fold cross validation for various configurations of the baseline and RNN systems on the training set. The results are presented in Table 1 and Table 2.

The best performing CRF system i.e. Baseline + Brown Clusters, achieves F1-score of 89.88. Except for brown clusters, additional features derived from distributional semantics, such as binarized word embeddings (BinEmb), prototype embeddings (ProtoEmb) contribute marginally to performance of the system. Pre-trained clinical embeddings improve F1 score by 11.93%, over random initialization of RNNs. In terms of recall, the RNN initialized with MIMIC embeddings is found to perform remarkably well without hand-engineered features. However, it fails to beat the CRF system at F1-score. Comparative analysis of individual entity classes reveals that the RNN improves recall for evidential and clinical department phrases by 5.44% and 8.32% respectively. It registers some drop in precision, but improves F1-score by up to 3%. Based on these results, we build the hybrid sequence tagger where the best performing CRF system is combined with RNN. The former is trained to tag problem, test and treatment and the latter is trained to tag rest of the three entity classes. The results are merged in a combination module and overlapping predictions are resolved by prioritizing the first three classes. The hybrid model improves recall by 2.36% and F1-score by 0.56% over the best-performing CRF system, when evaluated on test set.

With biomedical embeddings, we are able to develop a neural network which performs as good as CRF in within-domain clinical entity recognition for both, i2b2 and MTSamples datasets. Both CRF and RNN score beyond 94% and the difference in F1-score is 0.03%, 0.23% for i2b2 and MTSamples datasets respectively. While CRF systems have better precision, RNN systems achieve better recall and F1-score. Moreover, CRF systems are consistently out-performed by RNN systems in out-of-domain evaluation. CRF trained on MTSamples data scores 83.02% on i2b2 data. In a similar setting, RNN scores 1.23% higher than CRF. In a reverse setting, the CRF system achieves 86.54% F1 score, but is surpassed by more than 4% by the RNN system. The i2b2 dataset is almost double the size of MTSamples dataset and is probably more representative of MTSamples, than MTSamples is of the former. This could explain the wider gap in performance when systems are trained on i2b2 and tested on MTSamples. It is also eveident that, when semantics are borrowed from a larger corpus as in pre-trained word embeddings, it improves the system's domain adaptability. All the neural network architectures evaluated for domain adaptation also undergo fine-tuning of embeddings. If they are to be kept static as suggested by Tourille et al., performance might improve beyond the current results.

For intra-sentence event classification, the con-

System	Training set	Test set	TP	FP	FN	Precision	Recall	F1 Score
CRF	i2b2	i2b2	8912	279	856	**96.96**	91.24	94.01
RNN			8974	343	794	96.32	**91.87**	**94.04**
CRF	MTSamples	i2b2	7589	923	2179	89.16	77.68	83.02
RNN			7653	724	2115	**91.36**	**78.35**	**84.35**
CRF	MTSamples	MTSamples	9025	369	747	**96.07**	92.71	**94.36**
RNN			9170	582	602	94.03	**94.22**	94.13
CRF	i2b2	MTSamples	5556	713	960	89.61	83.67	86.54
RNN			9058	246	714	**88.77**	**93.10**	**90.89**

Table 6: Results for within-domain and out-of-domain performance of various systems and datasets

Task	Relation Class	TP	FP	FN	Precision	Recall	F1 Score
Relation Classification	Before	200	207	277	**49.14**	41.92	45.24
	After	97	345	208	**21.94**	31.80	25.96
	Overlap	1248	396	227	**75.91**	59.31	66.59
	Overall						68.45

Table 7: Results for temporal relation classification from i2b2 2012 dataset

Hyperparameter	Value
Hidden layers	1
Word embedding dimension	100
Position embedding dimension	5
Type embedding dimension	10
POS tag embedding dimension	5
Filter sizes	[2,3,4]
Number of filters	100

Table 8: Hyperparameters for CNN architecture

volutional network architecture scores 68.45%. Addition of negative instances (almost 40000 instances) makes the training set highly skewed and ill-suited for relation extraction. Preliminary results of relation extraction fare well below 50% for the positive relation classes.

5 Discussion

The hybrid architecture serves as a concept extraction model with a predisposition for higher recall of clinical events, as compared to the CRF system which exhibits better precision in performance. On comparing errors, we found the %overlap between false negatives of CRF and RNN systems to be only about 52%. The CRF model is able to exploit semantic, syntactic and orthographic information among others, while RNNs are only initialized with limited semantic information. Automatic learning of syntactic structure and finer semantic correlations is inherent to recurrent neural architecture. However, this may be somewhat limited by our small corpus. This situation leads to subtle disparities in performance of both systems.

The RNN is able to detect clinical departments (which includes physician names, hospitals names and clinical departments) with good recall value in spite of being trained with only 997 data points. CRF has lowest recall for clinical department, among all classes that contain more noun phrases. The RNN confuses higher percentage of *Treatment* concepts as *Occurrence* than CRF, mostly those which are verb phrases like 'excised', 'intubated' etc. Instead of initializing all words with clinical embeddings, the performance of RNN may be improved by selectively initializing clinical terms only. This can be done by filtering for certain UMLS semantic groups/types and providing only those words with a pre-trained word vector. On the other hand, word embeddings help the RNN in handling unseen vocabulary effectively. For example, when RNN is trained to tag 'decreased' as occurrence, it tags the word 'weaned' correctly as occurrence in the test set. Under similar conditions, CRF is unable to make the correct decision. Word vectors derived from a larger biomedical corpus may enable the RNN to make finer semantic distinctions. Unlike RNN, CRF fails to recognize the occasional long phrases such as *'normal appearing portal veins and hepatic arteries'*, even under overlap matching crite-

ria. We expect the LSTM cells in RNN to capture long-term dependencies from various ranges within a sentence, and our hypothesis is confirmed by the test results. The CRF operates within a prespecified context window and is limited by its linear chain framework. With a skip chain CRF, this situation can be remedied.

The percentage of words initialized by MIMIC and PubMed-PMC enbeddings is 78.96% and 66.84% respectively. In spite of the coverage being lower for the latter, it fares better than MIMIC embeddings. The biomedical embeddings are derived from a corpus that is larger as well as more inclusive of generic english words. This highlights the importance of high-quality semantic features in clinical NER tasks. Structured prediction methods fail to leverage this information, as we have seen from the lack of improvement in performance with addition of binarized or prototype embeddings (see Table 1). It also proves as a bottle-neck in domain adaptation. Not surprisingly, the best performing team in Clinical TempEval's domain adaptation task uses pre-trained embeddings and neural architectures. (Tourille et al.)

Convolutional networks have been proven to be efficient at capturing relations from free-text with minimum feature engineering. However, temporal relations might need more cues from long-term dependencies than can be captured by convolutional networks, which could be provided with LSTM cells. The first step to better performance for relation extraction is the availability of a less-skewed dataset, which may be solved with candidate-generation modules based on binary classifiers.

6 Conclusion & Future Work

In this work, we have performed a comparative analysis of CRF and RNN systems for clinical entity recognition. It can be concluded that, with high-quality word embeddings, RNNs achieve competitive performance and feature engineering is be minimized as a result. Through error analysis, we highlight some of the situations where RNNs fare better such as longer concept length, unseen clinical terms, semantically similar generic words, proper nouns etc. RNNs are also better suited for out-of-domain tasks because they make inference using semantics from a much larger corpus than the training data. We observe that clinical entity recognition systems trained on discharge summaries perform almost as well on other notes too, such as ER visit, progress notes, consult notes etc.

In future work, more variations in neural architecture will be explored for the extraction of temporal relations. We will also develop a module for the detection of temporal expressions from clinical text. Thereafter, it can be combined with the existing clinical entity recognition and relation extraction modules to form an end-to-end temporal evaluation system.

References

Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel M. Cer, and David Jurgens, editors. 2017. *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*. Association for Computational Linguistics. http://aclanthology.info/volumes/proceedings-of-the-11th-international-workshop-on-semantic-evaluation-semeval-2017.

Raghavendra Chalapathy, Ehsan Zare Borzeshi, and Massimo Piccardi. 2016. Bidirectional lstm-crf for clinical concept extraction. *arXiv preprint arXiv:1611.08373* .

Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. Neural temporal relation extraction. *EACL 2017* page 746.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting embedding features for simple semi-supervised learning. In *EMNLP*. pages 110–120.

Abhyuday N Jagannatha and Hong Yu. 2016a. Bidirectional rnn for medical event detection in electronic health records. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. NIH Public Access, volume 2016, page 473.

Abhyuday N Jagannatha and Hong Yu. 2016b. Structured prediction models for rnn based sequence labeling in clinical text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*. NIH Public Access, volume 2016, page 856.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data* 3.

Shantanu Kumar. 2017. A survey of deep learning methods for relation extraction. *arXiv preprint arXiv:1705.03645* .

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* .

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs). http://www.chokkan.org/software/crfsuite/.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics* 5:101–115.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.

Sunil Kumar Sahu, Ashish Anand, Krishnadev Oruganty, and Mahanandeeshwar Gattu. 2016. Relation extraction from clinical texts using domain invariant convolutional neural network. *arXiv preprint arXiv:1606.09370* .

L Smith, Thomas Rindflesch, W John Wilbur, et al. 2004. Medpost: a part-of-speech tagger for biomedical text. *Bioinformatics* 20(14):2320–2321.

Sunghwan Sohn, Kavishwar B Wagholikar, Dingcheng Li, Siddhartha R Jonnalagadda, Cui Tao, Ravikumar Komandur Elayavilli, and Hongfang Liu. 2013. Comprehensive temporal information detection from clinical text: medical events, time, and tlink identification. *Journal of the American Medical Informatics Association* 20(5):836–842.

Jannik Strötgen and Michael Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pages 321–324.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association* 20(5):806–813.

Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C Denny, and Hua Xu. 2013. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association* 20(5):828–835.

Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende, and Meliha Yetisgen-Yildiz. 2012. Statistical section segmentation in free-text clinical records. In *LREC*. pages 2001–2008.

Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. ???? Limsi-cot at semeval-2017 task 12: Neural architecture for temporal information extraction from clinical narratives .

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* 18(5):552–556.

Yonghui Wu, Jun Xu, Min Jiang, Yaoyun Zhang, and Hua Xu. 2015. A study of neural word embeddings for named entity recognition in clinical text. In *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, volume 2015, page 1326.

Yan Xu, Kai Hong, Junichi Tsujii, I Eric, and Chao Chang. 2012. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association* 19(5):824–832.

Yan Xu, Yining Wang, Tianren Liu, Junichi Tsujii, I Eric, and Chao Chang. 2013. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association* 20(5):849–858.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network. In *COLING*. pages 2335–2344.