Clinical Event Detection with Hybrid Neural Architecture

Adyasha Maharana
Biomedical and Health Informatics
University of Washington, Seattle
adyasha@uw.edu

Meliha Yetisgen
Biomedical and Health Informatics
University of Washington, Seattle
melihay@uw.edu

Abstract

Event detection from clinical notes has been traditionally solved with rule based and statistical natural language processing (NLP) approaches that require extensive domain knowledge and feature engineering. In this paper, we have explored the feasibility of approaching this task with recurrent neural networks, clinical word embeddings and introduced a hybrid architecture to improve detection for entities with smaller representation in the dataset. A comparative analysis is also done which reveals the complementary behavior of neural networks and conditional random fields in clinical entity detection.

1 Introduction

Event detection from clinical notes is a well studied problem in biomedical informatics; yet, it is constantly evolving through research. Much of this research has been promoted by the i2b2 challenges (2010, 2012) and their publicly available datasets comprised of annotated discharge summaries. For the 2010 task, the notes were annotated for three types of events - *Problem*, *Test* and *Treatment*, which are predominantly noun phrases. (Uzuner et al., 2011) The task was made even more challenging in 2012 with the addition of three new entity classes - *Occurrence*, *Evidential* and *Clinical Department*. *Occurrence* and *Evidential* concepts are mostly verb phrases, with some examples being 'readmitted', 'diagnosed', 'seen in consultation', 'revealed' etc. 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) 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 defined a CRF baseline to compare the performance of our neural networks and performed a detailed error analysis.

2 Related Work

The best performing system on 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. Most of the other competing teams also employed CRF for this task along with Support Vector Machines or Maximum Entropy for classifying the event category, with the exception of Jindal and Roth (2013) who implemented a sentence-level inference strategy using Integer Quadratic Program. 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 2012 i2b2 corpus has remained relatively unexplored in light of recent advances in NLP. We analyze the performance of recurrent neural networks for identification of clinical events from this dataset.

3 Methods

3.1 Dataset

The 2012 i2b2 corpus is made of 310 discharge summaries consisting of 178, 000 tokens annotated with clinical events, temporal expressions and temporal relations. The entire corpus is divided into training and test sets, containing 190 and 120 documents respectively. Each discharge summary has sections for clinical history and hospital course. 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.

3.2 Approach

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 (± 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).

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. Among others, its components include a char-RNN that models word prefixes, suffixes and shape - features that are critical to NER. We initialize two instances of the complete network with the GloVe vectors extracted from MIMIC-III discharge summaries. First instance is trained to classify problem, test and treatment concepts only; second instance is trained for other three classes. 78.96% words in the training corpus are initialized with pre-trained embeddings. 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.

3.2.4 Hybrid Architecture

The current of state-of-art for detecting problem, test and treatment concepts from clinical text is

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

based on CRF and it has been hard to improve on this baseline, even with neural networks. (Chalapathy et al., 2016) Cross-validation performance (presented in Table 2) reveals 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.

4 Evaluation Metrics and Results

We report the micro-averaged precision, recall and F1-score, for 'overlap' match of event spans as 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%. Clearly, RNNs are better suited for detecting occurrence, evidential and clinical department phrases from clinical text.

Based on these results on the training set, 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. We evaluate its performance on the i2b2 2012 test set. Results are listed in Table 3 and 4.

The hybrid model improves recall by 2.36% and F1-score by 0.56% over the best-performing CRF system. Dramatic improvement in recall (as high as 14%) is noted for some entities, but a similar drop in precision is observed.

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 3: 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 4: Entity-level performance of best performing CRF and Hybrid CRF-RNN on 2012 i2b2 test set

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 criteria. 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 pre-specified context window and is limited by its linear chain framework. With a skip chain CRF, this situation can be remedied.

6 Conclusion & Future Work

This paper evaluates various methods for using neural architecture in clinical entity recognition and minimizing feature engineering. Benefits are observed when the merits of structured prediction model and RNN are fused into a hybrid architecture after analysis of their cross-validation performance. The hybrid model's recall and F1 score surpass that of the state-of-art system we have used for replication. 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.

In future work, we will attempt to integrate long-term dependencies within a sentence by implementing the skip chain CRF model and explore the efficient use of word embeddings for structured prediction. This clinical entity recognition model will also be extended to a temporal evaluation system.

References

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

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.

Prateek Jindal and Dan Roth. 2013. Extraction of events and temporal expressions from clinical narratives. *Journal of biomedical informatics* 46:S13–S19.

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.

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

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

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

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

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.

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.