ATTENTION-BASED DEEP LEARNING SYSTEM FOR NEGATION AND ASSERTION DETECTION IN CLINICAL NOTES

Long Chen

Med Data Quest, Inc.
La Jolla, CA, USA
longchen@meddataquest.com

ABSTRACT

Natural language processing (NLP) has been recently used to extract clinical information from free text in Electronic Health Record (EHR). In clinical NLP one challenge is that the meaning of clinical entities is heavily affected by assertion modifiers such as negation, uncertain, hypothetical, experiencer and so on. Incorrect assertion assignment could cause inaccurate diagnosis of patients' condition or negatively influence following study like disease modelling. Thus, high-performance clinical NLP systems which can automatically detect negation and other assertion status of given target medical findings (e.g. disease, symptom) in clinical context are highly demanded. Here in this work, we propose a deep-learning system based on word embedding and Attention-based Bidirectional Long Short-Term Memory networks (Att-BiLSTM) for assertion detection in clinical notes. Unlike previous state-of-art methods which require knowledge input, our system is a knowledge poor machine learning system and can be easily extended or transferred to other domains. The evaluation of our system on public benchmarking corpora demonstrates that a knowledge poor deep-learning system can also achieve high performance for detecting negation and assertions comparing to state-of-the-art systems.

KEYWORDS

Natural language processing, Deep learning, BiLSTM, Clinical assertion, Attention

1. INTRODUCTION

A lot of valuable information are contained in clinical notes (e.g. patient medical history, discharge summaries, radiology reports and laboratory test results) of Electronic health records (EHRs) which can be used for various applications such as clinical decision support, disease modelling, medical risk evaluation, medication reconciliation, and quality measurements[1]. However, those clinical notes which are unstructured and in free text format, are difficult and time consuming for humans to manually review or analyse.

Therefore, Natural language processing (NLP) approaches have been developed for extracting useful information from clinical notes, such as medical concepts, patients' conditions and so on. However, accurate extraction of clinical entities is yet not enough as the real meaning of an extracted entity from clinical context is significantly affected by assertion modifiers such as negation, uncertainty, hypothetical, conditional, experiencer and so on. For instance, the entity "chest pain" in context "the patient denies chest pain" should be negated and not counted in the patient's condition. Previous studies[2] show that nearly half of the clinical concepts found in clinical notes are affected by assertion modifiers especially negation. Incorrect assertion assignment could cause inaccurate diagnosis of patients' condition, contaminate selected study cohorts or negatively influence following study such as disease modelling. Table 1 shows the assertion types we considered in this work as well as their examples. These assertion types are: absent (negation), hypothetical, possible (uncertainty), conditional (present in the patient under

DOI : 10.5121/ijaia.2019.10101

certain circumstances) and associated with someone else (AWSE). Moreover, unlike concept/entity extraction task, assertion detection requires both detecting the modifiers (e.g. trigger words) and deciding whether to assign the assertion relation/status to the target concept or not. Figure 1 shows one example. In Figure 1a, "No evidence of" is a valid negation modifier to the target clinical concept "other bowel pathology", so the target concept should be negated. However, in the same sentence as shown in Figure 1b, the current target concept "bright red blood per rectum" should not be affected by the negation trigger of "No evidence of". Because of these facts, assertion detection is one of the significant and challenging tasks in clinical NLP.

Here in this work, we propose using word embedding and Attention-based Bidirectional Long-Short Term Memory networks (Att-BiLSTM) for negation and assertion detection in clinical notes. We show that a knowledge-pool deep learning system based on Att-BiLSTM networks can also achieve good performance compared to state-of-art systems even with relatively small training dataset.

Table 1. Clinical assertions and examples

Assertion	**Concept**	**Context**
Absent (Negation)	splenomegaly	A follow-up CT scan was done which did not show any evidence for **splenomegaly** or hepatomegaly.
Hypothetical	infection	Dressing to remain in place for 10 days unless signs of bleeding, **infection** or is soiled.
Possible	bleeding	Given this, it was advised that the patient have a colonoscopy to rule out further **bleeding**.
Conditional	mild pain	Physical examination of the RLE showed **mild pain** in right hip with some movements.
AWSE	Breast cancer	**Breast cancer** in multiple female relatives.

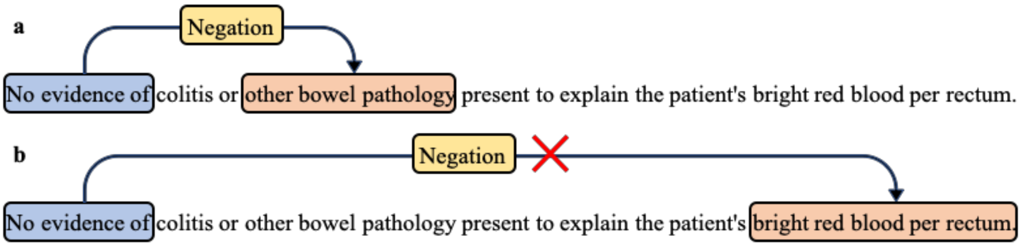
a
Negation
No evidence of
colitis or
other bowel pathology
present to explain the patient's bright red blood per rectum.
b
Negation
No evidence of
colitis or other bowel pathology present to explain the patient's
bright red blood per rectum.

Figure 1. Demonstration of assertion detection task. (a) a valid negation example; (b) an invalid negation example even though negation trigger presents

2. Related Work

Many previous works and NLP challenges contribute to address this issue, such as the 2010 Integrating Biology and the Bedside (i2b2)/Veteran's Affairs (VA) challenge (i2b2/VA) for assertion classification[3], the CoNLL-2010 shared task on hedges and their scope detection[4], and series of work focusing on negation detection[2],[5],[6]. Various approaches have been developed, and most of them can be classified as rule-based, machine learning or hybrid. The current widely used systems are mostly rule-based systems (e.g. NegEx, ConText, pyConTextNLP) which rely on manually generated rules using lexicon or syntax features such as trigger terms, termination clues, POS tag and dependency graph. Among them, NegEx[2] is one of the most popular and widely used system for negation. Here in this work, we also conducted focused study on negation using NegEx as the baseline compared to our models. Machine learning-based algorithms were also developed[3]. However, most of them still depend on knowledge input, feature engineering and using traditional classifiers such as support vector machine. Knowledge-poor system that purely relies on neural networks and deep learning is rare for assertion detection, which reflects the fact that there are not much public shared assertion annotated data in the community.

Attention-based neural network architectures which can help networks to selectively focus on particular information, recently gain much attention and have been proven to be effective in several NLP tasks such as machine translation[7] and relation classification[8]. The attention mechanism which was initially proposed as a solution for Encoder-Decoder model[7], is based on the idea that we need to select the most relevant information to compute the neural response, rather than using all available information. In NLP aspect, attention mechanism can help the model to focus on the words which have nontrivial effect on the target, and automatically capture the semantic information in a sentence, without using knowledge inputs such as lexicon or syntax patterns. With these recent advancements in deep learning research, we explored the possibility to apply attention-based bidirectional LSTM architecture for negation and assertion detection for clinical notes.

3. METHODS

In this section, we will describe the architecture of the model and the data we used for this study.

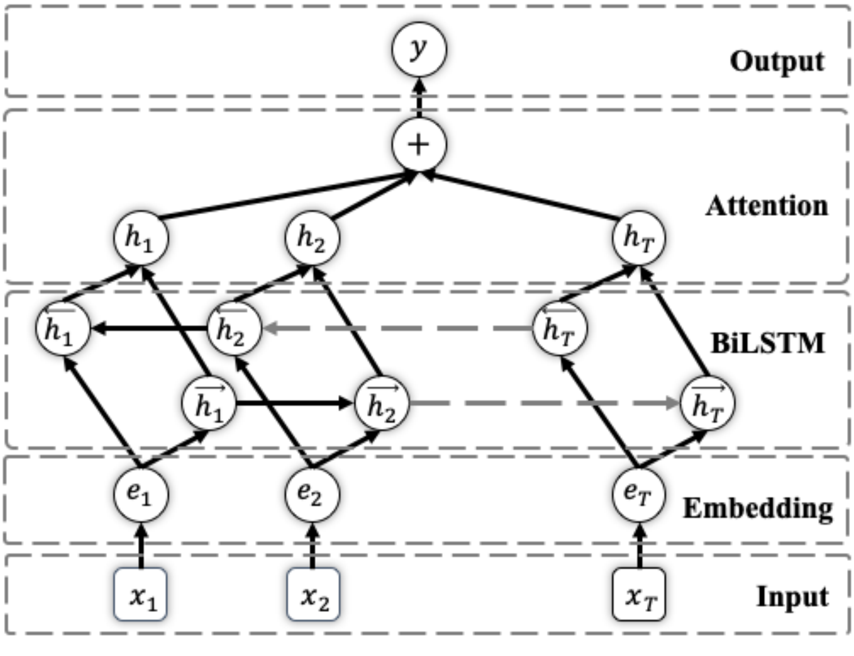

Figure 2. Structure of the attention-based bidirectional LSTM network

3.1. Model architecture

We used Att-BiLSTM architecture introduced by Zhou et al.[8] for assertion detection and assignment for given clinical concepts. We employed this approach combing attention mechanism and BiLSTM for two purposes. Firstly, considering the nature of this task which asks for detecting assertion status of target concept in sentence, attention mechanism can help the model to focus on the target concept and figure out the most important information related to the target. Secondly, LSTM is one of the most popular networks used for deep learning-based NLP and has been proven effective in various NLP tasks. Besides, in order to capture the sequence semantic information both forward and backward, we used bidirectional LSTM instead of unidirectional LSTM.

The architecture of this network is shown in Figure 2. As shown in Figure 2, the model contains five parts:

Input layer: The original context input of the model. Typically, this network takes positional marked concept and surrounding tokens as inputs. For instance, the sentence "The patient denies chest pain." will be prepared as "The patient denies <c> chest pain </c>.", where position markers are used to address the target concept.

Embedding layer: The input context is tokenized, and each word is mapped into a low dimension vector. Here in this study, we only considered word embedding and different home-trained or pre-trained word embeddings were tested. More details regarding each word embedding will be discussed in Section 4. In general, given a sentence $S = \{x_1, x_2, \dots, x_T\}$, for each word x_i can be projected to the whole vocabulary and treated as one-hot vector v_i with dimension of the total number of words in the vocabulary. Then a pre-trained word embedding can further transfer the one-hot vector to a low dimensional (200 as used in this study) real-valued vector:

$$e_i = W^{wrd} v^i \tag{1}$$

Then the sentence which is initially represented as a sequence of words is transferred as a sequence of numerical vectors.

LSTM layer: LSTM is used to obtain high level features containing temporal and syntax information. Here bidirectional LSTM was used in order to consider words before or after the target. LSTM networks typically contains tree components: input gate (i_t), forget gate (f_t), output gate (o_t). In between, hidden state (h_t) and cell state (c_t) as well as the corresponding weight matrix W and b serve to transfer the relations between gates along the sequence:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{2}$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right) \tag{3}$$

$$g_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + W_{cc}c_{t-1} + b_c) \tag{4}$$

$$c_t = i_t g_t + f_t c_{t-1} \tag{5}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o) \tag{6}$$

$$h_t = o_t tanh(c_t) \tag{8}$$

Where σ is the sigmoid function; x_t is the input vector at timestep t; i_t, f_t, o_t, h_t, c_t stand for the input gate, forget gate, output gate, hidden state and cell state respectively. W_{xi}, W_{xf}, W_{xc}, W_{xo} are the weight matrix of x_t on different gates respectively; W_{hi}, W_{hf}, W_{hc}, W_{ho} are the weight matrix of h_t on different gates; b_i, b_f, b_c, b_o are the bias offsets of the corresponding gates.

As shown in Figure 2, the sequence information propagates unidirectionally along the timesteps in the forward LSTM. Thus, only information in previous words and current word can be used to compute the neural response. In order to capture the information before and after the current word, both the forward LSTM and backward LSTM outputs were employed. Thus, the final output for each timestep (h_t) is generated by using element-wise sum of both forward and backward outputs:

$$h_t = [\overrightarrow{h_t} \oplus \overleftarrow{h_t}] \tag{9}$$

Where $\oplus$ is the elementwise plus. $\overrightarrow{h_t}$, $\overleftarrow{h_t}$ and h_t are the output of forward LSTM, output of backward LSTM and final output respectively.

Attention layer: The attention layer is designed to help the model recognize which part of the input data is important during the training. This layer allows the networks to selectively focus on specific information by generating a weight vector. After multiplying the weight vector, word-level features from each timestep are merged into sentence-level feature vector:

$$M = tanh(H) \tag{10}$$

$$\alpha = softmax(w^T M) \tag{11}$$

$$h^* = tanh(H\alpha^T) \tag{12}$$

Here H is the output matrix from LSTM layer; h^* is the final sentence level representation for assertion detection/classification; w is the weight matrix which is defined during training.

Output layer: Fully connected to target task and utilizes the sentence-level feature vector for assertion classification.

More details of this method can be found elsewhere[8].

2.2. Datasets

Assertion detection/classification was a subtask of the 2010 i2b2/VA NLP challenge. The corpus for this task along with the annotations is available for download. This data set includes patient discharge summaries and progress notes, which are the main data used for this research. There are 73 discharge summaries collected from Beth Israel Deaconess Medical Center, 97 from Partners HealthCare and 256 progress notes from University of Pittsburgh Medical Center. For the assertion annotated data, target concepts (medical problem) had assertion labels as either present, absent, possible, conditional (present in the patient under certain circumstances), hypothetical and associated with someone else (ASWE).

In addition, another dataset which is available in the NegEx source code was also used. This dataset contains concepts (clinical conditions) as well as corresponding sentences, extracted from 116 clinical notes at the University of Pittsburgh Medical Center. However, in this dataset only negation status has been annotated. Thus, concepts in sentences have been annotated either negated or affirmed.

Using the above corpora we constructed three datasets: 1) The dataset available with the NegEx rule-based system, referred to as the NegExCorp dataset; 2) The original i2b2 training dataset: i2b2 subsets from the Beth Israel Deaconess Medical Center and from Partners Healthcare, referred to as the i2b2-BID/PH dataset; and 3) The original 2010 i2b2 test dataset from University of Pittsburgh Medical Center, henceforth referred to as the i2b2-UPMC dataset. Table 2 provides more detailed information of these datasets.

Table 2. Distribution of assertions in each dataset

Datasets		**i2b2-BID/PH**	**i2b2-UPMC**	**NegEx-Corp**
Notes		170	256	116
Assertions	**Present**	4624	8622	1885
	Absent	1596	2594	491
	Hypothetical	382	445	-
	Possible	309	652	-
	Conditional	73	148	-
	AWSE	89	131	-
	Total	7073	12592	2376

4. Experimental Results

We implemented the Att-BiLSTM networks as described above and conducted cross datasets training/evaluation, in order to see the generalizability of this approach. For example, we trained the models on i2b2-BID/PH dataset and tested their performance on i2b2-UPMC and NegEx-Crop dataset. Table 3 shows the hyperparameters we used during training. Evaluation was conducted under three metrics:

Assertion: This is the general assertion evaluation which takes all six assertion categories into consideration. Here for each category, data in that class were treated as positive and all data in other classes were treated as negative.

Neg1: In order to compare with NegEx, "present" class was treated as positive and all other classes were treated as negative. This metric is targeted at evaluating system performance of recognizing all kind of reasons which indicates not presented on the patient.

Neg2: In order to compare with NegEx, "present" class was treated as positive and only "absent" as negative, ignoring data in other categories. This metric is targeted at evaluating system performance on detecting only the negation (e.g. "no evidence of", "denies") status of the concept.

We evaluated our systems using precision, recall and F1 score on five-fold cross validation. Micro-F1 score was used as our main evaluation metrics for Assertion evaluation as following the 2010 i2b2/VA challenge. For Neg1 and Neg2 evaluations, the F1 score of "Negated" class was used as the main metric. We compared the performance of our models against NegEx rule-based system as baseline when focusing on negation.

Table 3. Network hyperparameters

Parameters	**Value**
Word dimension	200
LSTM unit size	128
Dropout	0.5
Regularization	1e-4
Learning rate	1e-4

4.1. Testing on word embeddings

Two word embeddings as well as a control of using random initialized word vector as the baseline was tested: 1) Random: randomly generated 200-dimensional word vectors for the vocabulary in training dataset; 2) MIMIC: a home developed word embedding trained on MIMIC III datasets[9] with word2vec algorithm; 3) PubMed+: a word embedding trained by Jagannatha et al.[10] with PubMed Open Access articles, an unlabelled EHR corpus and the English Wikipedia. Table 4 shows the performance (F1 score on each assertion class) of Att-BiLSTM models with different word embeddings. Here in this experiment, models were trained on i2b2-BID/PH datasets and conducted assertion evaluation on i2b2-UPMC dataset. As shown in Table 4, using word embedding can improve the model in every classes and the PubMed+ word embedding showed the greatest improvement.

Table 4. F1-score of the models trained with different word embeddings

	Word Embedding		
	Random	**MIMIC**	**Pub-Med+**
Present	0.939	0.940	**0.950**
Absent	0.912	0.921	**0.927**
Hypothetical	0.830	0.846	**0.865**
Possible	0.499	0.557	**0.637**
Conditional	0.435	0.459	**0.544**
AWSE	0.729	0.743	**0.743**
Ave. (micro)	0.899	0.905	**0.922**

Table 5. Three evaluation metrics on model trained with PubMed+ word embedding

Eval. Metrics		**Pre.**	**Rec.**	**F1**
Assertion	**Present**	0.938	0.962	0.950
	Absent	0.923	0.931	0.927
	Hypothetical	0.857	0.874	0.865
	Possible	0.772	0.541	0.637
	Conditional	0.667	0.460	0.544
	AWSE	0.798	0.695	0.743
	Ave. (micro)	0.918	0.922	0.922
Neg1	**Affirmed**	0.938	0.962	0.950
	Negated	0.912	0.861	0.886
	Ave. (micro)	0.930	0.930	0.930
Neg2	**Affirmed**	0.983	0.962	0.973
	Negated	0.882	0.946	0.913
	Ave. (micro)	0.960	0.958	0.959

4.2. Evaluation on single model

In order to have a detailed evaluation of the model on assertion and a focused view on negation as compared to NegEx, three evaluation metrics were used as mentioned above. Table 5 shows the evaluation results on i2b2-UPMC dataset of the model trained on i2b2-BID/PH dataset with

PubMed+ word embedding. And Table 5 also shows the model performance (Precision, Recall and F1 scores) on negation. A detailed comparison of this model with NegEx also conducted under Neg1 and Neg2 evaluation metrics on the Negated class (Table 6). As shown in Table 5, model achieved an overall micro-F1 score of 0.922 for assertion detection and assignment task. This high score is comparable with the state-of-art systems as reported in 2010 i2b2/VA challenge on assertion classification where the micro-F1 scores of the top 10 systems range from 0.921 to 0.936[3]. Considering that the model requires no feature engineering or domain knowledge input, this result indicates that a knowledge-poor neural network based deep learning system can also achieve high performance in assertion task even with a relatively small training dataset. The focused view on negation task (Table 6) shows that the Att-BiLSTM based model overperformed NegEx on i2b2-UPMC data under both Neg1 and Neg2 evaluation metrics, which further demonstrates the capability of this approach.

Table 6. Model vs NegEx performance on negation with i2b2-UPMC data

Systems		Pre.	Rec.	F1
Neg1	**NegEx-Neg.**	0.907	0.545	0.681
	Model-Neg.	**0.912**	**0.861**	**0.886**
Neg2	**NegEx-Neg.**	**0.903**	0.793	0.845
	Model-Neg.	0.882	**0.946**	**0.913**

Table 7. Cross datasets evaluation under Assertion, Neg1 and Neg2 evaluation metrics

		Test Datasets		
Systems		**i2b2-BID/PH**	**i2b2-UPMC**	**NegEx-Corp**
Assertion	**Model trained on:**			
	i2b2-BID/PH	-	0.922	-
	i2b2-UPMC	0.931	-	-
Neg1	**NegEx**	0.696	0.681	0.929
	Model trained on:			
	i2b2-BID/PH	-	**0.886**	0.926
	i2b2-UPMC	**0.911**	-	**0.931**
Neg2	**NegEx**	0.864	0.845	0.929
	Model trained on:			
	i2b2-BID/PH	-	**0.913**	0.926
	i2b2-UPMC	**0.932**	-	**0.931**

4.3. Evaluation crossing datasets

In order to test the generalizability of this approach, we did more detailed cross datasets evaluation. We trained the Att-BiLSTM models on i2b2-BID/PH or i2b2-UPMC datasets separately and evaluate them on each other's dataset as well as the dataset attached with NegEx source code: NegEx-Corp. Table 7 shows the results on evaluating models or NegEx on different datasets under Assertion, Neg1 and Neg2 evaluation metrics. Here in this table, the micro-F1 score was used for Assertion evaluation and the F1 score of "Negated" class was used for Neg1 and Neg2 evaluation. Table 7 shows that the Att-BiLSTM models trained on i2b2-BID/PH or i2b2-UPMC datasets performance at same level with slightly difference on evaluating each other's datasets or NegEx-Corp. For negation task, both of the Att-BiLSTM models obviously outperform NegEx on i2b2-dataset and show similar level of performance on NegEx-Corp. These results demonstrate the generalizability of this approach crossing different datasets or institutes. A comparation among the Att-BiLSTM models shows that the model trained on i2b2-UPMC performs better than the other, which indicates that further benefit may be expected if trained with larger datasets.

5. CONCLUSIONS

In this work, we introduce a deep learning system based on word embedding and attention-based bidirectional LSTM networks for automatically assertion detection and assignment in clinical notes. We also conducted cross datasets evaluation on public benchmarking corpora for assertion classification as well as focused study on negation. The evaluation in comparison with other state-of-the-art systems demonstrates the capability and generatability of this approach. Our results indicate that a knowledge poor deep learning system can also achieve high performance for detecting assertions and compares favourably to state-of-the-art systems.

REFERENCES

[1] Demner-Fushman, D., Chapman, W.W. and McDonald, CJ. (2009) "What can natural language processing do for clinical decision support?", J Biomed Inform., 42(5):760–772.

[2] Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G. (2001). "A simple algorithm for identifying negated findings and diseases in discharge summaries.", J Biomed Inform., 34:301–10.

[3] Uzuner, O., South, B., Shen, S., and DuVall, S.L. (2011) "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text.", J Am Med Inform Assoc., 18:552-556.

[4] Farkas, R., Vincze, V., Móra, G., Csirik, J., and Szarvas, G., (2010) "The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text.", Proceedings of the Fourteenth Conference on Computational Natural Language Learning - Shared Task, p.1-12, July 15-16, 2010, Uppsala, Sweden.

[5] Harkema, H., Dowling, J.N., Thornblade, T., Chapman, W.W., (2009) "ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports.", J Biomed Inform., 42:839–51.

[6] Mehrabi, S., Krishnan, A., Sohn, S., Roch, A.M., Schmidt, H., Kesterson, J., Beesley, C., Dexter, P., Schmidt, C.M., Liu, H., and Palakal, M., (2015) "DEEPEN: a negation detection system for clinical text incorporating dependency relation into NegEx." J Biomed Inform., 54:213–9.

[7] Bahdanau, D., Cho, K. and Bengio. Y., (2014) "Neural machine translation by jointly learning to align and translate.", arXiv preprint arXiv:1409.0473.

[8] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H. and Xu, B., (2016) "Attention-based bidirectional long short-term memory networks for relation classification.", In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, volume 2, pages 207–212

[9] Johnson, A., Pollard, T., Shen, L. and Lehman, L. (2016) "MIMIC-III, a freely accessible critical care database.", Scientific Data, 3,160035.

[10] Jagannatha, A.N. and Yu, H., (2016) "Bidirectional rnn for medical event detection in electronic health records.", In Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting, volume 2016, page 473.