# Joint Entity Extraction and Assertion Detection for Clinical Text

**Parminder Bhatia**
Amazon, USA
parmib@amazon.com

**Busra Celikkaya**
Amazon, USA
busrac@amazon.com

**Mohammed Khalilia**
Amazon, USA
khallia@amazon.com

## Abstract

Negative medical findings are prevalent in clinical reports, yet discriminating them from positive findings remains a challenging task for information extraction. Most of the existing systems treat this task as a pipeline of two separate tasks, i.e., named entity recognition (NER) and rule-based negation detection. We consider this as a multi-task problem and present a novel end-to-end neural model to jointly extract entities and negations. We extend a standard hierarchical encoder-decoder NER model and first adopt a shared encoder followed by separate decoders for the two tasks. This architecture performs considerably better than the previous rule-based and machine learning-based systems. To overcome the problem of increased parameter size especially for low-resource settings, we propose the *Conditional Softmax Shared Decoder* architecture which achieves state-of-art results for NER and negation detection on the 2010 i2b2/VA challenge dataset and a proprietary de-identified clinical dataset.

## 1 Introduction

In recent years, natural language processing (NLP) techniques have demonstrated increasing effectiveness in clinical text mining. Electronic health record (EHR) narratives, e.g., discharge summaries and progress notes contain a wealth of medically relevant information such as diagnosis information and adverse drug events. Automatic extraction of such information and representation of clinical knowledge in standardized formats (Singh and Bhatia, 2019) could be employed for a variety of purposes such as clinical event surveillance, decision support (Jin et al., 2018), pharmacovigilance, and drug efficacy studies.

Although many NLP applications that successfully extract findings from medical reports have been developed in recent years, identifying assertions such as positive (present), negative (absent), and hypothetical remains a challenging task, especially to generalize (Wu et al., 2014). However, identifying assertions is critical since negative and uncertain findings are frequent in clinical notes (Figure 1), and information extraction algorithms that do not distinguish between them will not paint a clear picture of the patient.

> Discontinue Abraxane, patient denies taking Tyleno 325 mg and is not taking calcium carbonate. Patient also stopped taking colecalciferol 1,000 units PO.

Figure 1: Negated medications (highlighted in red) and negation cues (highlighted in purple) in clinical text. Our model does not explicitly label the cues.

In this paper, we focus on identifying the negated findings in a multi-task setting (Bhatia et al., 2018). Most of the existing systems treat this task as a pipeline of two separate tasks, i.e., named entity recognition (NER) and negation detection. Previous efforts in this area include both rule-based and machine-learning approaches.

Rule-based systems rely on negation keywords and rules to determine the cue of negation. NegEx (Chapman et al., 2001) is a widely used algorithm that consists of ontology lookup to index findings, and negation regular expression search in a fixed scope. ConText (Harkema et al., 2009) extends NegEx to other attributes like hypothetical and make scope variable by searching for a termination term. NegBio (Peng et al., 2018) uses a universal dependency graph for scope detection. Another similar work by Gkotsis et al. (2016) utilizes a constituency-based parse tree to prune out the parts outside the scope. However, these approaches use rules and regular expressions for cue detection which rely solely on surface text



*Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 954–959
Florence, Italy, July 28 - August 2, 2019. 

and thus are limited when attempting to capture complex syntactic constructions such as long noun phrases.

Kernel-based approaches are also very common, especially in the 2010 i2b2/VA task of predicting assertions. The state-of-the-art in that challenge applies support vector machines (SVM) to assertion prediction as a separate step after concept extraction (de Bruijn et al., 2011). They train classifiers to predict assertions of each concept word, and a separate classifier to predict the assertion of the whole concept. Shivade et al. (2015) propose an Augmented Bag of Words Kernel (ABoW), which generates features based on NegEx rules along with bag-of-words features. Cheng et al. (2017) use CRF for classification of cues and scope detection. These machine learning based approaches often suffer in generalizability, the ability to perform well on unseen text.

Recently, neural network models by Fancellu et al. (2016) and Rumeng et al. (2017) have been proposed. Most relevant to our work is that of Rumeng et al. (2017) where gated recurrent units (GRU) are used to represent the clinical events and their context, along with an attention mechanism. Given a text annotated with events, it classifies the presence and period of the events. However, this approach is not end-to-end as it does not predict the events. Additionally, these models generally require large annotated corpus, which is necessary for good performance. Unfortunately, such clinical text data is not easily available.

Multi-task learning (MTL) is one of the most effective solutions for knowledge transfer across tasks. In the context of neural network architectures, we perform MTL by sharing parameters across models, such as pretraining using word embeddings (Bhatia et al., 2016; Bojanowski et al., 2016), a popular approach for most NLP tasks. In this paper, we propose an MTL approach to negation detection that overcomes some of the limitations in the existing models such as data accessibility. MTL leverages overlapping representation across sub-tasks and it is one of the most effective solutions for knowledge transfer across tasks. In the context of neural network architectures, we perform MTL by sharing parameters across tasks.

To the best of our knowledge, this is the first work to jointly model named entity and negation in an end-to-end system. Our main contributions are summarized below:

- An end-to-end hierarchical neural model consisting of a shared encoder and different decoding schemes to jointly extract entities and negations. Using our proposed model, we obtain substantial improvement over prior models for both entities and negations on the 2010 i2b2/VA challenge task as well as a proprietary de-identified clinical note dataset for medical conditions.
- *A Conditional softmax shared decoder* model to overcome low resource settings (datasets with limited amounts of training data), which achieves state of art results across different corpora.
- A thorough empirical analysis of parameter sharings for low resource setting highlighting the significance of the shared decoder.

## 2 Methodology

We first present a standard neural framework for named entity recognition. To facilitate multi-task learning, we expand on that architecture by building a two decoder model. Then, to overcome the issues of the two decoder model we propose a single shared decoder model. Finally, we introduce the *Conditional softmax shared decoder*.

### 2.1 Named Entity Recognition Architecture

NER is a sequence tagging problem which maximizes a conditional probability of tags $\mathbf{y}$ given an input sequence $\mathbf{x}$, parameterized by $\theta$.

$$P(\mathbf{y}|\mathbf{x};\theta) = \prod_{t=1}^{T} P(y_t|x_t, y_{<t};\theta) \quad (1)$$

Here $T$ is the length of the sequence, and $y_{<t}$ represents tags for all previous time-steps. We focus on an established hierarchical architecture (Lample et al., 2016; Yang et al., 2016; Chiu and Nichols, 2016) consisting of encoders (at both word and character levels) and a tagger for output generation.

#### 2.1.1 Encoders

Input to the model, $\mathbf{x} \in \mathbb{N}^T$, represents token ids of the input vocabulary. This sequence is encoded first at the character level and additionally at the word level. Character level representation consists of using a bi-directional Long Short-Term Memory (BiLSTM) (Hochreiter and Schmidhuber, 1997; Graves et al., 2013) unit to encode each

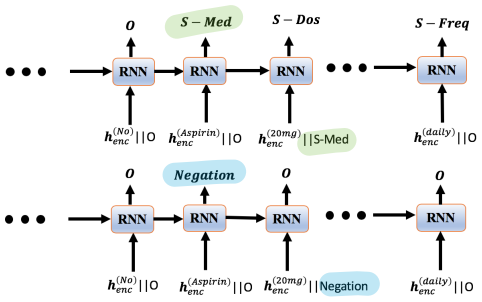


Figure 2: Two decoder model, upper decoder for NER and the lower decoder for negation, where common encoder

word independently. For each word we subsequently have sequences $\overrightarrow{h^{(t)}_{1:l}}$ and $\overleftarrow{h^{(t)}_{1:l}}$, where $l$ represents the length of the word. We concatenate the last time-step of each of these sequences to obtain a vector representation, $h^{(t)}_c = [\overrightarrow{h^{(t)}_l}||\overleftarrow{h^{(t)}_l}]$. The final input to the word level encoder is a combination of a pre-trained word embedding (Pennington et al., 2014) and the character representation, $m_t = [h^{(t)}_c||\text{emb}_{word}(x_t)]$. For the word level encoder we make use of another BiLSTM.

#### 2.1.2 Tagger

The tagger consists of a uni-directional LSTM which takes as input the latent word representation given by the word level encoder, as well as the label embedding of the previously generated tag. During training we feed ground truth labels by way of teacher forcing (Williams and Zipser, 1989), while at test time we use the generated sequence directly. This system is trained using a standard cross-entropy objective.

### 2.2 Two Decoder Model

To facilitate the MTL setting, we begin with a two decoder model consisting of decoders which use the shared encoder representation to jointly predict entities and negation attribute (Figure 2). This is a standard architecture used for MTL which consists of different LSTM's for decoders followed by corresponding softmaxes. This model mitigates the issues associated with rule-based models that rely solely on surface text, and thus are limited when attempting to capture complex syntactic constructions. With shared contextual encoder representation consisting of character and word embedding based models, the proposed architecture provides an effective solution for knowledge transfer across tasks, thus consolidating the ability to perform well on unseen text. However, this proposed architecture is not scalable, the number of decoders scales linearly with the number of attributes. Another problem we realized with this architecture is the performance degradation when working in an extremely low resource setting, where more parameters prevent the model from generalizing well.

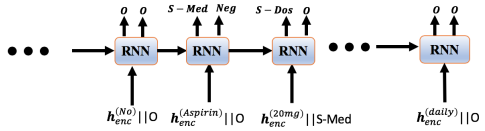


Figure 3: Shared decoder model

### 2.3 Shared Decoder Model

To overcome the limitations of the two decoder model we propose a shared decoder model (Figure 3). We share the encoder and decoder of the two tasks and the common output from the decoder is fed into two different softmax for entity and negations.

$$\hat{y}^{Entity}_t = \text{Softmax}^{Ent}(\mathbf{W^{Ent}}o_t + b^s)$$
$$\hat{y}^{Neg}_t = \text{Softmax}^{Neg}(\mathbf{W^{Neg}}o_t + b^s)$$

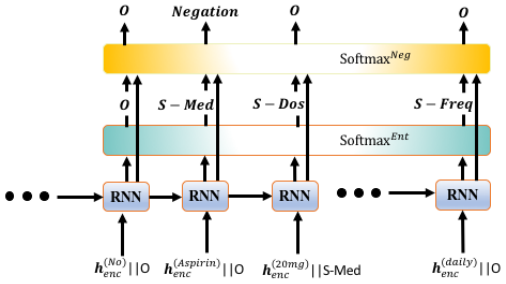


Figure 4: Conditional softmax decoder model

#### 2.3.1 Conditional Softmax Decoder Model

While the single decoder model is more scalable, we found that this model did not perform as well for negation as the two decoder model. It can be attributed to the fact that negation occurs less frequently than the entities, thus the decoder primarily focuses on making entity extraction predictions. To mitigate this issue and provide more context to negation attributes, we add an additional input, which is the softmax output from entity extraction (Figure 4). Thus, the model learns more about the input as well as the label distribution from entity extraction prediction. As an example, we use negation only for PROBLEM entity in the

| | | 2010 i2b2/VA | | | Proprietary Med. Cond. | | |
|---|---|---|---|---|---|---|---|
| | Model | P | R | $F_1$ | P | R | $F_1$ |
| NER | LSTM:CRF (R. Chalapathy and Piccardi, 2016) | 0.844 | 0.834 | 0.839 | 0.820 | 0.840 | 0.830 |
| | Independent NER (Lample et al., 2016) | 0.857 | 0.841 | 0.848 | 0.880 | 0.848 | 0.863 |
| | Two Decoder (this paper) | 0.849 | 0.855 | 0.851 | 0.876 | 0.861 | 0.868 |
| | Shared Decoder (this paper) | 0.852 | 0.821 | 0.834 | 0.864 | 0.841 | 0.85 |
| | **Conditional** (this paper) | 0.854 | 0.858 | **0.855** | 0.878 | 0.872 | **0.874** |
| NEGATION | Negex (Chapman et al., 2001) | 0.896 | 0.799 | 0.845 | 0.403 | 0.932 | 0.563 |
| | ABoW Kernel (Shivade et al., 2015) | 0.899 | 0.900 | 0.900 | - | - | - |
| | Independent Negation (Lample et al., 2016) | 0.810 | 0.850 | 0.820 | 0.840 | 0.820 | 0.83 |
| | Two Decoder (this paper) | 0.894 | 0.908 | 0.899 | 0.931 | 0.865 | 0.897 |
| | Shared Decoder (this paper) | 0.870 | 0.902 | 0.882 | 0.921 | 0.850 | 0.878 |
| | **Conditional** (this paper) | 0.919 | 0.891 | **0.905** | 0.928 | 0.874 | **0.899** |

Table 1: Test set performance during multi-task training. The table displays precision, recall and macro averaged $F_1$. The baseline is the current state-of-the art optimized architecture.

i2b2 dataset. Providing the entity prediction distribution helps the negation model to make better predictions. The negation model learns that if the prediction probability is not inclined towards PROBLEM, then it should not predict negation irrespective of the word representation.

$$\hat{y}_t^{Ent}, \text{SoftOut}_t^{Ent} = \text{Softmax}^{Ent}(\mathbf{W^{Ent}} o_t + b^s)$$
$$\hat{y}_t^{Neg} = \text{Softmax}^{Neg}(\mathbf{W^{Neg}}[o_t, \text{SoftOut}_t^{Ent}] + b^s)$$

where, $\text{SoftOut}_t^{Ent}$ is the softmax output of the entity at time step $t$.

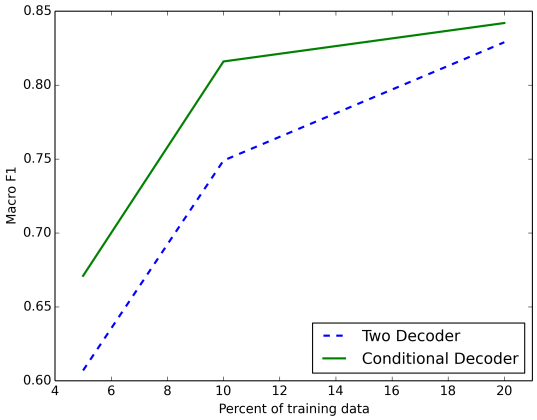


Figure 5: Conditional softmax decoder is more robust in extreme low resource setting than its two decoder counterpart.

## 3 Experiments

### 3.1 Dataset

We evaluated our model on two datasets. First is the 2010 i2b2/VA challenge dataset for "test, treatment, problem" (TTP) entity extraction and assertion detection (*i2b2 dataset*). Unfortunately, only part of this dataset was made public after the challenge, therefore we cannot directly compare with NegEx and ABoW results. We followed the original data split from R. Chalapathy and Piccardi (2016) of 170 notes for training and 256 for testing. The second dataset is proprietary and consists of 4,200 de-identified, annotated clinical notes with medical conditions (*proprietary dataset*).

### 3.2 Model settings

Word, character and tag embeddings are 100, 25, and 50 dimensions, respectively. For word embeddings we use GloVe (Peng et al., 2018) and fine tune during training, while character and tag embeddings are randomly initialized. Character and word encoders have 50, and 100 hidden units, respectively, while the tagger LSTM has a hidden size of 50. Dropout is used after every RNN, as well as for word embedding input. We use Adam (Kingma and Ba, 2014) as an optimizer. Hyperparameters are tuned using Bayesian Optimization (Snoek et al., 2012).

## 4 Results

Since there is no prior work which has solved the two tasks as a joint model, we report the best results for both the individual tasks (Table 1). We observe that the baseline model for NER (**Indepedent NER**) presented in the methodology section outperforms the best model (R. Chalapathy and Piccardi, 2016) on the i2b2 challenge. The **Two decoder** and the conditional softmax decoder (**Conditional decoder**) model achieve even better results for NER than our baseline model, where

the conditional decoder model achieved new state-of-art for 2010 i2b2/VA challenge task. **Shared decoder** underperformed the other two models. That can be attributed to a single decoder which primarily focuses on making entity extraction predictions which are more frequent than negations. The conditional decoder outperformed the baseline model on the negation prediction task and achieved an improvement of about $8\%$ in $F_1$ score compared to the baseline model, which suggests that modeling named entity and negation tasks together helps in achieving better results than each of the tasks done independently.

We compare our models for negation detection against NegEx, and ABoW which has best results for the negation detection task on i2b2 dataset. Conditional decoder model outperforms both NegEx and ABoW (Table 1). Low performance of NegEx and ABoW is mainly attributed to the fact that they use ontology lookup to index findings and negation regular expression search within a fixed scope. A similar trend was observed in the medication condition dataset. The important thing to note is the low $F_1$ score for NegEx. This can primarily be attributed to abbreviations and misspellings in clinical notes which can not be handled well by rule-based systems.

To understand the advantage of conditional decoder, we evaluated our model in extreme low data settings where we used a sample of our training data. We observed that the conditional decoder outperforms the two decoder model and achieved an improvement of $6\%$ in $F_1$ score in those settings (Figure 5). As we increase the data size, their performance gap narrows in demonstrating that the conditional decoder is robust in low resource settings.

## 5 Conclusion

In this paper we have shown that named entity and negation assertion can be modeled in a multi-task setting. Joint learning with shared parameters provides better contextual representation and helps in alleviating problems associated with using neural networks for negation detection, thereby achieving better results than the rule-based systems. Our proposed conditional softmax decoder achieves best results across both tasks and is robust to work well in extreme low data settings. For future work, we plan to investigate the model on other related tasks such as relation extraction, normalization as well as the use of advanced conditional models.

## References


Parminder Bhatia, Kristjan Arumae, and Busra Celikkaya. 2018. Dynamic transfer learning for named entity recognition. *arXiv preprint arXiv:1812.05288*.

Parminder Bhatia, Robert Guthrie, and Jacob Eisenstein. 2016. Morphological priors for probabilistic neural word embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 490–500.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Berry de Bruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 2011. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5):557–562.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Katherine Cheng, Timothy Baldwin, and Karin Verspoor. 2017. Automatic negation and speculation detection in veterinary clinical text. In *Proceedings of the Australasian Language Technology Association Workshop 2017*, pages 70–78.

Jason Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association of Computational Linguistics*, 4(1):357–370.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2016. Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 495–504.

George Gkotsis, Sumithra Velupillai, Anika Oellrich, Harry Dean, Maria Liakata, and Rina Dutta. 2016. Don't let notes be misunderstood: A negation detection method for assessing risk of suicide in mental health records. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 95–105.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE.

Henk Harkema, John N Dowling, Tyler Thornblade, and Wendy W Chapman. 2009. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42(5):839–851.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Mengqi Jin, Mohammad Taha Bahadori, Aaron Colak, Parminder Bhatia, Busra Celikkaya, Ram Bhakta, Selvan Senthivel, Mohammed Khalilia, Daniel Navarro, Borui Zhang, et al. 2018. Improving hospital mortality prediction with medical named entities and multimodal learning. *arXiv preprint arXiv:1811.12276*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*, pages 260–270.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2017:188.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

E. Z. Borzeshi R. Chalapathy and M. Piccardi. 2016. Bidirectional lstm-crf for clinical concept extraction. *arXiv preprint arXiv:1611.08373*.

Li Rumeng, N Jagannatha Abhyuday, and Yu Hong. 2017. A hybrid neural network model for joint prediction of presence and period assertions of medical events in clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1149. American Medical Informatics Association.

Chaitanya Shivade, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M Lai. 2015. Extending negex with kernel methods for negation detection in clinical text. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, pages 41–46.

Gaurav Singh and Parminder Bhatia. 2019. Relation extraction using explicit context conditioning. *arXiv preprint arXiv:1902.09271*.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. 2014. Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PloS one*, 9(11):e112774.

Zhilin Yang, Ruslan Salakhutdinov, and William Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *arXiv preprint arXiv:1603.06270*.