Remedi: A Medical Information Extraction System

Valerija Holomjova
Department of Artificial Intelligence
Faculty of Information Communication Technology
University of Malta
valerija.holomjova.17@um.edu.mt

ABSTRACT

This research presents a biomedical information extraction tool called Remedi, specialized in identifying 'medical problem' and 'treatment' entities from unstructured text in the medical domain, as well as the relations between them. The tool consists of a Biomedical Named Entity Recognition (BM-NER) model which employs a Bidirectional-Long Short-Term Memory Conditional Random Field (Bi-LSTM-CRF) model, as well as a Biomedical Relation Extraction (BM-RE) model which consists of a Bi-LSTM model. A subset of the i2b2 2010 challenge dataset has been acquired to train and evaluate the BM-NER and BM-RE components. We evaluated the performance of the Bi-LSTM and Bi-LSTM-CRF models when given additional external features such as the UMLS and the GENIA tagger. Results showed that with the addition of external features, both models outperformed similar models that did not leverage on such features. The BM-NER system achieved a micro-average F1 score of 0.846 which can be ranked second amongst top-performing models presented in the i2b2 2010 concept extraction challenge. On the other hand, the BM-RE system achieved a micro-average F1 score of 0.652 and needs further improvement to reach the performance of the best models in the i2b2 2010 relation extraction challenge.

Keywords

Natural Language Processing, Biomedical Named Entity Recognition, Biomedical Relation Extraction, Information Extraction, Bi-LSTM, Bi-LSTM-CRF, CRF, Machine Learning, Deep Learning

1. INTRODUCTION

Medical knowledge can often be found in narrative form such as medical textbooks, clinical records or online published medical reports. Natural Language Processing (NLP) is a branch of Artificial Intelligence (AI) which can be used to extract meaningful information from textual data. Numerous medical reports are being published online daily on websites such as PubMed[1], containing information on advancements in the medical industry. Automated information extraction tools [1, 2, 3] have been used to infer new structured knowledge from such sources with the scope of supporting both researchers and medical students. Information extraction tools rely heavily on annotated corpora, however, in the case of Electronic Medical Records (EMR), corpora are scarce since these documents contain sensitive patient information, including medical history and prescribed treatments [4]. As a result, advancements in the creation of tools that extract information from EMRs has been slow [1, 2, 3].

There have been a number of competitions [5, 6, 7] that involved NLP tasks related to the medical domain. These competitions consisted of tasks such as identifying medical entities and relations from text, determining temporal relations in clinical narratives and identifying medical risk factors, such as those related to Coronary Artery Disease (CAD), in medical records. Popular approaches used to address such problems employed machine learning models [8, 9, 10, 11], however, more recent research [12, 13, 14] suggests that Recurrent Neural Networks (RNN) have the potential of outperforming them. An LSTM is a popular RNN architecture that is capable of learning long term dependencies. Bi-LSTM-CRF models are a variety of LSTM based models and have shown promising results in sequential tagging tasks [15]. Research has also shown that the addition of features from external tools [16, 17] can improve the performance of the developed models [8, 18, 19].

This research will investigate research that focused on the extraction of entities and relations from biomedical text such as clinical reports. Furthermore, we present a system called Remedi, that is targeted at identifying medical problem and treatment entities, and extracting medical problem-treatment and medical problem-medical problem relations from unstructured biomedical text. This research will also investigate the performance of a Bi-LSTM-CRF and Bi-LSTM model, with additional features from external tools [16, 17], on BM-NER and BM-RE tasks respectively. The two hypotheses that we want to address through our research are; "the addition of external features will increase the performance of a Bi-LSTM model for BM-RE tasks" and "the addition of external features will increase the performance of a Bi-LSTM-CRF model for BM-NER tasks".

1.1 Aims and Objectives

The main aim behind this research is to extract medical entities and medical problem-treatment and medical problem-medical problem relations from unstructured biomedical text. The following research question will be addressed in this research;

> "Can medical entities and relations associated with medical problems and treatments be extracted from unstructured biomedical text?"

This research question will be tackled by fulfilling the following objectives;

[1] https://www.ncbi.nlm.nih.gov/pubmed/, accessed on 12/05/2020

- **Objective 1 (O1):** We will research and develop a BM-NER tool that is able to locate and distinguish between problem and treatment entities in unstructured biomedical data. We will compare different models and evaluate whether the performance of the BM-NER will improve with the addition of features from external tools [16, 17], and then compare it to a constructed baseline model. Hyperparameter tuning will be carried out to improve the results of the model.
- **Objective 2 (O2)**: We will research and develop a BM-RE solution that identifies relations between medical problem and treatment entities. Similar to O1, the performance of the model with the addition of external features will be evaluated and compared to baseline models implemented in other research [12]. Additionally, hyperparameter tuning will be carried out to maximize the performance of the model.

2. BACKGROUND AND LITERATURE REVIEW

2.1 Background

Named Entity Recognition (NER) is a sub-task of information extraction that focuses on extracting concepts from unstructured text such as names or locations. Biomedical Named-Entity Recognition (BM-NER) focuses on the recognition of medical entities such as genes, proteins [20, 21] or disease names [7]. Early-based BM-NER systems were rule-based [1, 2, 3] and relied on medical vocabularies such as the Unified Medical Language System (UMLS) [16]. The UMLS is a collection of controlled vocabularies of biomedical terms that can be used to add further context to words in unstructured biomedical text. In recent research, tools employ Machine Learning (ML) techniques [8, 9, 22], specifically CRFs which are discriminative classifiers that use contextual information from neighbouring labels to predict sequences.

Relation Extraction (RE) is the task of extracting semantic relations between two or more entities. These systems are typically pattern-based [23] or ML-based [10, 18]. Pattern-based approaches use hand-coded rules which require the involvement of domain experts. ML-based approaches are dependent on annotated data and model relation extraction as a classification problem. Popular ML approaches use Support Vector Machines (SVM) [10, 11, 24] which are models that plot features in a multidimensional space and differentiates their class labels through the construction of hyperplanes. This research focuses on the identification of two main class of relations from the i2b2 challenge [7]; treatment-problem relations and problem-problem relations.

Novel research has explored the use of RNNs in BM-NER and BM-RE tasks [12, 13, 14]. RNNs are a type of Artificial Neural Networks (ANN) which make use of sequential information. The unique feature of RNN's is that their connections form a direct cycle, hence the layers of the neural network are dependent on each other. In other words, the output of previous layers can be passed as input to the next layer. LSTM is a popular RNN architecture that is capable of learning long term dependencies. Bi-LSTM-CRF models have been used for sequence tagging [15] and consists of training two LSTM's on an input sequence and adding a CRF layer for sequential labelling. The key features of the model are that the two LSTM layers are provided with the same input sequence in the model but as reversed copies, and the CRF layer is then used to predict the best label sequence.

2.2 Literature Review

In 2010, the i2b2/VA Workshop [7] presented an NLP challenge consisting of three tasks; concept extraction, assertion classification, and relation extraction of medical entities using an annotated corpus of de-identified clinical reports.

2.2.1 Biomedical Named Entity Recognition

During the concept extraction task of the i2b2 2010 challenge [7], participants were tasked with the extraction of three medical entities from the provided dataset. The three entities to be extracted were related to medical problems, treatments and tests. A majority of the participants of the challenge that achieved the best results [8, 9, 22, 25] tackled the task using CRF models. De Bruijn et al. [19] were an exception as they did not use the CRF approach, but created a concept extraction system that generated the highest F-measure result in the challenge for the concept extraction task (0.852), by using Semi Markov Models (SMM) trained using an online learning algorithm referred to as the Passive-Aggressive (PA) algorithm. The participants of the challenge [7] demonstrated that systems achieved higher results using rich textual features and the inclusion of knowledge-rich sources such as the UMLS [8, 19, 22].

Boag et al. [25] creates CliNER, a two-pass machine learning system that uses a CRF classifier to establish concept boundaries and an SVM classifier for establishing concept types. Zhang and Elhadad [26] and Ghiasvand and Kate [27] present an unsupervised solution to BM-NER, which employs a signature-based approach. The unsupervised systems [26, 27] generated the lowest score compared to other systems, outlining the importance of having annotated corpora for BM-NER tasks.

Wu, Yonghui et al. examined the impact of word embeddings and deduced that they can be used as features to improve the performance of ML models for NER [28, 13]. They explored the performance of DL models using word embeddings on Chinese Clinical notes [28] and then the i2b2 2010 dataset [13]. They discovered that the RNN was able to outperform all other systems, achieving a high F-measure score of 0.859 [13]. Liu, Zengjian et al. [14] investigated the performance of an LSTM on the different datasets of the i2b2 challenge [5, 6, 29] and deduced that their system can outperform CRF models. Chalapathy et al. [30] explored the performance of a Bi-LSTM-CRF model using the i2b2 2010 dataset and experimented with various word embeddings and hyperparameters. Their model achieved a higher F1 score than most of the CRF models in the 2010 i2b2 challenge [9, 22], and ranked closely with top-ranked systems [8, 19]. Overall, it should be highlighted that most systems that used DL techniques [13, 14] outperformed the other systems using the same dataset [7].

2.2.2 Biomedical Relation Extraction

In the third task of the i2b2 2010 challenge [7], participants were asked to identify relations between the three entity types that were recognized in the concept extraction task. The relations could be categorized into three main categories; medical problem-treatment, medical problem-test and medical problem-medical problem relations.

Rink et al. [11] approached the task by implementing an SVM using rich sets of lexical and context features. Their system obtained the highest F-score of the relation extraction challenge [7] (0.737). Overall, the third task of the i2b2 challenge illustrated that most effective relation extraction systems adopted the SVM approach [11, 24, 31]. It was discovered that the provided corpus had several concept pairs that had no relations, hence, certain systems chose to remove those relations before identifying relations [19]. It should be noted that certain systems benefited from down-sampling and the augmentation of hand-built rules [24]. Most importantly, the papers highlighted that the performance of their system [31] could have significantly improved if provided with more labelled examples, indicating the need for domain knowledge in the annotation of these examples.

Frunza and Inkpen developed a supervised SVM-based classifier [10] and experimented with informative representations using the i2b2 2010 dataset [7]. Their system produced an F-measure score higher than the average score of the participants of the challenge (0.862). Hassan, Mohsen et al. [32] propose a novel method known as SPARE, which learns patterns from dependency graphs of given text which have at least one symptom and disease. The system is able to discover new relationships as well as new and complex symptoms. Chikka and Karlapalem [12] develop a Hybrid Bi-LSTM and rule-based system for medical relation extraction that was trained on a subset of the i2b2 2010 challenge [7]. The authors compared and contrasted the performance of an SVM and a Bi-directional LSTM (Bi-LSTM) to their developed system. The study concludes that their developed system obtains the highest F-score (0.52).

In conclusion, existing research has highlighted the limitation in performance due to the availability of a small corpus [31]. Overall, the majority of the top-performing systems employed machine learning supervised approaches and the use of the SVM model was the most popular method that produced some of the highest scores [10, 11, 24, 31]. Certain research outlined the importance of the use of rich features for the training of the ML models [11, 19]. A few papers also mentioned that the exploration of implementing knowledge bases such as UMLS could generate higher results for their system [18, 11].

3. METHODOLOGY

The design of the proposed information extraction system, Remedi, is shown in Figure 1. The user will provide some text to the system in the form of a sentence (or multiple sentences) and the system will output all detected entities in the given sentences, as well as the relations between any of the identified entity pairs. The sentences can be passed into the system in the form of an input text file. The system will then tokenize the sentences and carry out feature extraction. Two external tools (the GENIA tagger [17] and the UMLS [16]) will be used to extract additional features.

The tokenized text and extracted features will then be passed into the BM-NER model that will tag each word in the sentence as a medical problem entity ('B-problem' or 'I-problem'), treatment entity ('B-treatment' or 'I-treatment') or an entity that corresponds to neither of the two tags ('O'). The tags acquired from the BM-NER component will then be used to identify entity pairs within the sentences. These entity pairs are used to extract further features from the sentence, specifically, the distance between each entity in the entity pair and the other words in the sentence. Next, the entities acquired from the BM-NER and the entity distance features will then be passed with the previous features into the BM-RE model to classify the relations between each entity pair.

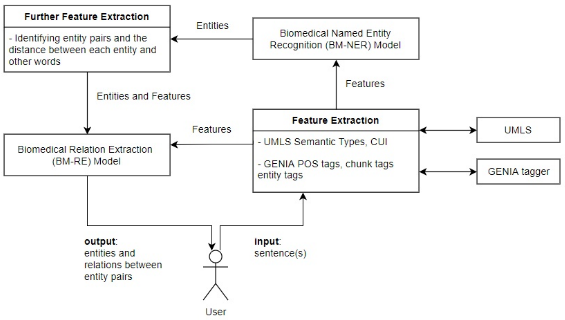
Further Feature Extraction
- Identifying entity pairs and the distance between each entity and other words
Entities
Biomedical Named Entity Recognition (BM-NER) Model
Features
Entities and Features
Features
Feature Extraction
- UMLS Semantic Types, CUI
- GENIA POS tags, chunk tags entity tags
UMLS
GENIA tagger
Biomedical Relation Extraction (BM-RE) Model
output:
entities and relations between entity pairs
input:
sentence(s)
User

Figure 1: Data-flow diagram for Remedi

The BM-RE model identifies two relation types; medical problem-medical problem relations and treatment-medical problem relations. The first class of relations is composed of two types of relations; medical problems that indicate other medical problems (PIP) and relations where the two medical problems entities do not associate with one another (NPPR). The second class of relations consists of six possible relations; treatments improving medical problems (TrIP), treatments worsening medical problems (TrWP), treatments causing medical problems (TrCP), treatments being administered for medical problems (TrAP), treatments not administered as a cause of a medical problem (TrNAP) and relations where the treatment and medical problem entities do not associate with one another (NTPR).

3.1 Biomedical Named Entity Recognition

The BM-NER system uses the following training features; POS tags, chunk tags and entity tags from the GENIA tagger, as well as the semantic types and CUI of each word from the UMLS. These features were used to train both the CRF model and the Bi-LSTM-CRF model. However, since the CRF model is more flexible with the addition of features, more features were extracted from the text, such as the POS tags from NLTK library [33]. General features such as the previous and following three unigrams, as well as features from the previous and following tokens (such as capitalization or whether the word is positioned at the start or end of the sentence) were used to train the CRF model, similar to the approach adopted by Boag et al. [25]. Each line in the medical reports is iterated through, trimmed and tokenized into a list of words which is tagged in Inside-Outside-Beginning (IOB) format using the annotated concepts from the dataset. It should be noted that the set of informative sentences includes sentences that have at least one tagged entity in them, whereas the set of non-informative sentences includes those that have no tagged entities.

CRFsuite [34] was used to construct the CRF model then evaluate its performance using standard Scikit-learn functions. To ensure that the best parameters were chosen GridSearchCV found in Scikit-learn was used for hyperparameter tuning. We explored the performance of the model using four different algorithms ('lbfgs', 'l2sgd', 'pa', 'arow') and

Table 1: Dataset distribution for BM-RE and BM-NER

<table>
<tr><th>Component</th><th colspan="10">Frequency</th></tr>
<tr><td rowspan="2">BM-NER</td><td colspan="3">Problem Entities
('B-problem', 'I-problem')</td><td colspan="3">Treatment Entities
('B-treatment', 'I-treatment')</td><td colspan="2">Untagged Entities
('O')</td><td>Informative Sentences</td><td>Non-Informative Sentences</td></tr>
<tr><td colspan="3">19,664</td><td colspan="3">14,185</td><td colspan="2">342,949</td><td>18,579</td><td>25,362</td></tr>
<tr><td rowspan="2">BM-RE</td><td>PIP</td><td>TrAP</td><td>TrCP</td><td>TrIP</td><td>TrNAP</td><td>TrWP</td><td>NTPR</td><td>NPPR</td><td>Informative Sentences</td><td>Non-Informative Sentences</td></tr>
<tr><td>2,203</td><td>2,616</td><td>526</td><td>203</td><td>174</td><td>133</td><td>4,462</td><td>27,215</td><td>5,855</td><td>31,677</td></tr>
</table>

their corresponding parameters, which were differed slightly from their default values using the following website[2].

The Bi-LSTM-CRF model was built using Keras and the evaluation of the model was carried out using Seqeval[3]. The model consists of six layers. The first set of layers are input layers, representing the BM-NER features of each sentence that are fed into the model, followed by their own embedding layers. All input features had to be vectorized by converting them into a sequence of integers where each integer is the index of a token in a feature vocabulary. The embedding layers are then concatenated and passed into the Bi-LSTM layer, which is followed by a time distributed dense layer and the final CRF layer, which predicts the best possible tagged sequence for a given sentence. During hyperparameter tuning we explored various activation functions ('relu', 'elu'), dropout and recurrent dropout rates (0.1, 0.2, 0.3), batch sizes (32, 64, 128), embedding sizes (50, 100, 150, 200, 250, 300) and optimizers ('rmsprop', 'adam'). Each model was trained on an epoch size of 5. A CRF loss function was used, and the accuracy of the CRF classifier, as well as the model, were used as an evaluation metric. Both the loss function and evaluation metric are part of Keras-contrib[4].

3.2 Biomedical Relation Extraction

The BM-RE system uses the same features that were used to train the Bi-LSTM-CRF model for the BM-NER component. Additionally, the BM-RE system uses the entities extracted from the BM-NER component as a training feature, and also to determine possible entity pairs within sentences. Moreover, the BM-RE considers the distance between an entity in an entity pair and each word of a sentence as a feature. The set of informative sentences for BM-RE includes sentences where entity pairs have a correlation between them ('PIP', 'TrAP', 'TrCP', 'TrIP', 'TrNAP', 'TrWP'), whereas the set of non-informative sentences includes sentences where entity pairs have no correlation between them ('NTPR' and 'NPPR').

The Bi-LSTM that was built for the BM-RE component was constructed in a similar manner to that built for the BM-NER, however, unlike the Bi-LSTM-CRF, the Bi-LSTM excludes the time distributed and CRF layer, as it will output a single classification for each sentence rather than a sequence. The evaluation of the model was carried out using Scikit-learn functions.

For the hyperparameter tuning of the Bi-LSTM, the model explores the same parameters as the Bi-LSTM-CRF model, however, instead of using the exponential linear unit ('elu') activation function, the softmax activation function was used. Each model was trained using an epochs value of 5. The Bi-LSTM model uses categorical cross-entropy as a loss function and accuracy as an evaluation metric.

4. EVALUATION

A subset of the i2b2 2010 challenge [7] was used for training and evaluating the BM-NER and BM-RE models. Table 1 depicts the distribution of the dataset that was used. For both components, the set of informative and non-informative sentences are shuffled and split for testing and evaluation purposes into a ratio of 1:4, to form a testing and training set respectively. The BM-NER dataset was well-distributed in contrast the BM-RE dataset which had an imbalance of relation annotations between medical problem-treatment and treatment-treatment relations.

The evaluation for both components of the system was carried out using micro-average precision, recall and F1 score. It should be noted that this is only the case during hyperparameter tuning of the components, as the final evaluation of the components considers both the micro-average and macro-average scores.

During hyperparameter tuning of the BM-NER components, it was discovered that that the best performing CRF model used the PA algorithm with a *c* value of 2 and a *pa_type* of 2. In the case of the Bi-LSTM-CRF, the best performing model used a word embedding size (WES) of 250, batch size (BS) of 32, recurrent dropout rate (RD) of 0.2, dropout rate (D) of 0.1, 'rmsprop' as the optimization (O) function and 'elu' as an activation (A) function. It was discovered that the Bi-LSTM-CRF models which generated the lowest results had the highest batch sizes (128) and used the 'adam' optimizer. On the other hand, models using smaller batch sizes, larger WES, and the 'rmsprop' optimizer generated higher results. The varying dropout rates, recurrent dropout rates and different activation functions had a generally even distribution amongst high and low results and did not impact the ranking as much.

During hyperparameter tuning of the Bi-LSTM model for the BM-RE component of Remedi, we discovered that the best performing model had a WES of 300, BS of 32, RD rate of 0.1, D rate of 0.1 and used the 'adam' optimizer and 'softmax' activation function. Furthermore, the models which generated the lowest results used the 'relu' activation function. On the other hand, models using a smaller BS, larger WES and the 'softmax' activation function generated higher results. Additionally, the TrAP relation was the highest detected relation in all the models generated during hyperparameter tuning, followed by the PIP and TrCP relations. The models had the hardest time detecting TrIP, TrNAP and TrWP relations.

Tables 2 and 3 compare the performance of the models we constructed for BM-NER and BM-RE to similar research. In relation to BM-NER, it was observed that systems obtained higher scores during the detection of 'treatment' entities in

[2] https://sklearn-crfsuite.readthedocs.io/en/latest/api.html#module-sklearn_crfsuite, accessed on 07/05/2020

[3] https://github.com/chakki-works/seqeval, accessed on 07/05/2020

[4] https://github.com/keras-team/keras-contrib, accessed on 14/05/2020

Table 2: Evaluation of the final model performance for BM-NER

System/ Authors	Model Type	Problem Precision (3 d.p.)	Problem Recall (3 d.p.)	Problem F1 score (3 d.p.)	Treatment Precision (3 d.p.)	Treatment Recall (3 d.p.)	Treatment F1 score (3 d.p.)	Evaluation Metric	Overall Precision (3 d.p.)	Overall Recall (3 d.p.)	Overall F1 score (3 d.p.)
CliNER [25]	CRF	0.710	0.858	0.777	0.834	0.752	0.791	Micro	0.795	0.812	0.800
Remedi	CRF	0.878	0.835	0.856	0.866	0.812	0.838	Micro	0.873	0.824	0.848
								Macro	0.872	0.823	0.846
Chalapathy et al. [30]	Bi-LSTM-CRF	-	-	-	-	-	-	Micro	-	-	0.838
Remedi	Bi-LSTM-CRF	0.865	0.825	0.844	0.883	0.818	0.849	Micro	0.873	0.822	0.846
								Macro	0.873	0.822	0.847

Table 3: Evaluation of the final model performance for BM-RE

System/ Authors	Model Type	Metric Type	PIP	TrAP	TrCP	TrIP	TrNAP	TrWP	Overall Score (Micro)	Overall Score (Macro)
Chikka and Karlapalem [12]	Bi-LSTM and Rule-based	Precision (2 d.p.)	-	0.61	0.47	0.44	0.22	0.15	0.51	-
		Recall (2 d.p.)	-	0.56	0.47	0.41	0.46	0.21	0.53	-
		F1 score (2 d.p.)	-	0.58	0.47	0.42	0.30	0.17	0.52	-
Remedi	Bi-LSTM	Precision (3 d.p.)	0.577	0.762	0.560	0.500	0.833	0.333	0.678	0.594
		Recall (3 d.p.)	0.497	0.762	0.560	0.333	0.588	0.333	0.627	0.512
		F1 score (3 d.p.)	0.534	0.762	0.560	0.400	0.690	0.333	0.652	0.546

comparison to ‘problem’ entities. Remedi’s baseline CRF model used the same library and similar features to CliNER [25], however, it managed to obtain higher overall precision, recall and F1 score than CliNER for both its macro-average and micro-average scores. Remedi’s Bi-LSTM-CRF model obtained a higher macro and micro F1 score in comparison to the model by Chalapathy et al. [30], highlighting the positive effect of external features on the performance of the Bi-LSTM-CRF model. On the other hand, the Bi-LSTM-CRF model did not outperform the baseline CRF model. Remedi’s Bi-LSTM was trained on the same subset of the 2010 i2b2 dataset as the hybrid Bi-LSTM model by Chikka and Karlapalem [12], and managed to obtain a higher micro-average overall precision, recall and F1 score than their system. These results highlight that the additional external features did improve the results of the Bi-LSTM model. Our Bi-LSTM model also outperformed the baseline SVM model and Bi-LSTM model implemented in their research [12], which obtained micro F1 scores of 0.46 and 0.51 respectively. .

5. CONCLUSION AND FUTURE WORK

We addressed our research question by presenting Remedi, a system tasked with identifying medical entities in unstructured text and the relations between them. Remedi consists of two components; a BM-NER and a BM-RE system, which were both trained and evaluated using a subset of the i2b2 2010 dataset [7].

We accomplished objective O1 by creating a Bi-LSTM-CRF model for the BM-NER component of Remedi that achieved a micro F1 score of 0.846, which can be ranked amongst the top-performing models in the i2b2 2010 concept extraction challenge. Objective O2 was accomplished by implementing a Bi-LSTM model for the BM-RE component of Remedi, which achieved a micro F1 score of 0.652. We believe that with further improvement, the model could achieve higher results especially if given more training data.

The performance of the Bi-LSTM and Bi-LSTM-CRF models both improved when given features from the UMLS and the GENIA tagger, and outperformed similar models that did not use these features, allowing us to prove our initial hypotheses as true. The Bi-LSTM-CRF model did not outperform the baseline CRF model, however, the Bi-LSTM model was able to outperform the baseline SVM model. The close difference in the results between the Bi-LSTM-CRF and CRF model suggests that adding more features could improve the performance of the Bi-LSTM-CRF model and allow it to outperform the CRF model.

For future work, we would like to investigate whether the performance of the BM-NER and BM-RE models will improve when given a larger corpus for training, such as the full version of i2b2 2010 dataset. Furthermore, it would be interesting to investigate whether our models will improve with the addition of more training feature, such as those used in CTAKES [35], or by adding pre-trained word embeddings from a larger corpus, such as the MIMIC-II dataset [36]. We would also like to develop a tool that uses Remedi to automatically infer and model structured knowledge from online medical articles found on sites like PubMed. Moreover, we would like to further investigate how the UMLS could be exploited to determine whether a problem concept represents a sign or symptom to be used for structuring knowledge in future tools.

6. REFERENCES

[1] A. R. Aronson, “Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.,” in *Proceedings of the AMIA Symposium*, p. 17, American Medical Informatics Association, 2001.

[2] J. C. Denny, P. R. Irani, F. H. Wehbe, J. D. Smithers, and A. Spickard III, “The KnowledgeMap project: development of a concept-based medical school curriculum database,” in *AMIA Annual Symposium Proceedings*, vol. 2003, p. 195, American Medical Informatics Association, 2003.

[3] C. Friedman, G. Hripcsak, W. DuMouchel, S. B. Johnson, and P. D. Clayton, “Natural language processing in an operational clinical information system,” *Natural Language Engineering*, vol. 1, no. 1, pp. 83–108, 1995.

[4] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, “Extracting information from textual documents in the electronic health record: a review of recent research,” *Yearbook of medical informatics*, vol. 17, no. 01, pp. 128–144, 2008.

[5] A. Stubbs, C. Kotfila, H. Xu, and Ö. Uzuner, “Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track

2," *Journal of biomedical informatics*, vol. 58, pp. S67–S77, 2015.

[6] W. Sun, A. Rumshisky, and O. Uzuner, "Evaluating temporal relations in clinical text: 2012 i2b2 Challenge," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 806–813, 2013.

[7] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 552–556, 2011.

[8] M. Jiang, Y. Chen, M. Liu, S. T. Rosenbloom, S. Mani, J. C. Denny, and H. Xu, "A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 601–606, 2011.

[9] N. Kang, R. J. Barendse, Z. Afzal, B. Singh, M. J. Schuemie, E. M. van Mulligen, and J. A. Kors, "Erasmus MC approaches to the i2b2 Challenge," in *Proceedings of the 2010 i2b2/VA workshop on challenges in natural language processing for clinical data*, i2b2, Boston, MA, USA, 2010.

[10] O. Frunza and D. Inkpen, "Extracting relations between diseases, treatments, and tests from clinical data," in *Canadian Conference on Artificial Intelligence*, pp. 140–145, Springer, 2011.

[11] B. Rink, S. Harabagiu, and K. Roberts, "Automatic extraction of relations between medical concepts in clinical texts," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 594–600, 2011.

[12] V. R. Chikka and K. Karlapalem, "A hybrid deep learning approach for medical relation extraction," *arXiv preprint arXiv:1806.11189*, 2018.

[13] Y. Wu, M. Jiang, J. Xu, D. Zhi, and H. Xu, "Clinical Named Entity Recognition Using Deep Learning Models," in *AMIA Annual Symposium Proceedings*, vol. 2017, p. 1812, American Medical Informatics Association, 2017.

[14] Z. Liu, M. Yang, X. Wang, Q. Chen, B. Tang, Z. Wang, and H. Xu, "Entity recognition from clinical texts via recurrent neural network," *BMC medical informatics and decision making*, vol. 17, no. 2, p. 67, 2017.

[15] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[16] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[17] Y. Tsuruoka, "GENIA tagger: Part-of-speech tagging, shallow parsing, and named entity recognition for biomedical text," *Available at: www-tsujii. is. su-tokyo. ac. jp/GENIA/tagger*, 2006.

[18] X. Lv, Y. Guan, J. Yang, and J. Wu, "Clinical relation extraction with deep learning," *International Journal of Hybrid Information Technology*, vol. 9, no. 7, pp. 237–248, 2016.

[19] B. de Bruijn, C. Cherry, S. Kiritchenko, J. Martin, and X. Zhu, "NRC at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features," in *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data. Boston, MA, USA: i2b2*, 2010.

[20] B. Settles, "ABNER: An open source tool for automatically tagging genes, proteins, and other entity names in text," *Bioinformatics*, vol. 21, no. 14, pp. 3191–3192, 2005.

[21] M. Torii, Z. Hu, C. H. Wu, and H. Liu, "BioTagger-GM: a gene/protein name recognition system," *Journal of the American Medical Informatics Association*, vol. 16, no. 2, pp. 247–255, 2009.

[22] M. Torii, K. Wagholikar, and H. Liu, "Using machine learning for concept extraction on clinical documents from multiple data sources," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 580–587, 2011.

[23] A. B. Abacha and P. Zweigenbaum, "Automatic extraction of semantic relations between medical entities: a rule based approach," *Journal of biomedical semantics*, vol. 2, no. 5, p. S4, 2011.

[24] C. Grouin, A. B. Abacha, D. Bernhard, B. Cartoni, L. Deleger, B. Grau, A.-L. Ligozat, A.-L. Minard, S. Rosset, and P. Zweigenbaum, "CARAMBA: concept, assertion, and relation annotation using machine-learning based approaches," 2010.

[25] W. Boag, K. Wacome, T. Naumann, and A. Rumshisky, "CliNER: A lightweight tool for clinical named entity recognition," *AMIA Joint Summits on Clinical Research Informatics (poster)*, 2015.

[26] S. Zhang and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts," *Journal of biomedical informatics*, vol. 46, no. 6, pp. 1088–1098, 2013.

[27] O. Ghiasvand and R. J. Kate, "Learning for clinical named entity recognition without manual annotations," *Informatics in Medicine Unlocked*, vol. 13, pp. 122–127, 2018.

[28] Y. Wu, M. Jiang, J. Lei, and H. Xu, "Named entity recognition in Chinese clinical text using deep neural network," *Studies in health technology and informatics*, vol. 216, p. 624, 2015.

[29] A. Stubbs and Ö. Uzuner, "Annotating risk factors for heart disease in clinical narratives for diabetic patients," *Journal of biomedical informatics*, vol. 58, pp. S78–S91, 2015.

[30] R. Chalapathy, E. Z. Borzeshi, and M. Piccardi, "Bidirectional LSTM-CRF for clinical concept extraction," *arXiv preprint arXiv:1611.08373*, 2016.

[31] J. D. Patrick, D. H. Nguyen, Y. Wang, and M. Li, "A knowledge discovery and reuse pipeline for information extraction in clinical notes," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 574–579, 2011.

[32] M. Hassan, O. Makkaoui, A. Coulet, and Y. Toussaint, "Extracting disease-symptom relationships by learning syntactic patterns from dependency graphs," 2015.

[33] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[34] N. Okazaki, "CRFsuite: a fast implementation of Conditional Random Fields (CRFs)," 2007.

[35] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[36] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database," *Critical care medicine*, vol. 39, no. 5, p. 952, 2011.