Azerbaijan Journal of High Performance Computing, Vol 4, Issue 2, 2021, pp. 232-241
https://doi.org/10.32010/26166127.2021.4.2.232.241

End-to-End Relation Extraction on Clinical Text Data using Natural Language Processing

Naveen S Pagad[1], Pradeep N[2]

[1] Visvesvaraya Technological University, Belagavi, Karnataka, India, naveenspagad@gmail.com
[2] Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India, nmnpradeep@gmail.com

*Correspondence: Naveen S Pagad, Visvesvaraya Technological University, Belagavi, Karnataka, India, naveenspagad@gmail.com

Abstract

In light of the increasing number of clinical narratives, a modern framework for assessing patient histories and carrying out clinical research has been developed. As a consequence of using existing approaches, the process for recognizing clinical entities and extracting relations from clinical narratives was subsequently error propagated. Thus, we propose an end-to-end clinical relation extraction model in this paper. Clinical XLNet has been used as the base model to address the discrepancy issue, and the proposed work has been tested with the N2C2 corpus.

Keywords: Clinical entity, Relation extraction, Error propagation, End-to-end model

1. Introduction

Data is the collection of facts and statistics concerning an object or originated an event. The processed data is information whose objective is to increase its usefulness of the data. Knowledge represents an understanding of certain information. The unstructured nature of text data makes automated understanding difficult and has led to the development of several text mining (TM) techniques in the last decade (Bose, P., et al., 2021).

In the clinical context, clinical data is raw data in a patient's Electronic Health Record (EHR). The clinical narrative written by a healthcare practitioner originated in the occurrence of an event, such as admission reports written during patient admission, lab reports, and discharge summaries. Clinical data from electronic health records are used for computerized clinical applications (e.g., clinical decision support systems) and clinical and translational research. EHR data can serve as a challenge because patient information can be embedded in the clinical text, which is not directly accessible through other computerized applications that utilize structured data (Tang, B., Cao, H., Wu, Y., Jiang, M., & Xu, H., 2013, April).

The clinical terms found and extracted from the clinical data, such as medication or diseases, is clinical information. Clinical knowledge comprises comprehension of the clinical data extracted, such as establishing clinical relations between the patient diagnosis and the clinical terms found in the patient's EHR. An example of clinical knowledge could be discovering which medications are more prescribed in a given clinical specialty based on the clinical information extracted from the EHRs' clinical data, written in narratives.

NLP technologies have been introduced in the medical domain in the past decade that extracts structured clinical information from narrative text (Friedman, C., et al., 1994). The term Natural Language Processing (NLP) refers to the methods used by computers to interpret spoken or written information. Several high-level tasks are needed to process human languages with NLP, such as machine translation, question-answering systems, information extraction, and natural language understanding. As part of data analysis, KDD, and data mining, knowledge extraction is essential to extracting structured information from unstructured data. A significant amount of continuous medical information is acquired, maintained, and available digitally, including background information, blood tests, medications, therapies, and therapeutic interventions.

Clinical Named Entity Recognition (NER) is an essential step in extracting patient data from medical records. A need for clinical NER in the mining of clinical data has attracted research attention. Specifically, it aims to build a database of unique medical entities from medical texts, where the target entities include diseases, medications, examinations, and remedies 9 Zhang, R., Zhao, P., Guo, W., Wang, R., & Lu, W., 2022). However, medical NER differs from generic NER in several important ways. There are many alternative spellings and synonyms that contribute to the development of vocabulary. As a result, medicine interpretation is negatively affected.

This research aims to extract end-to-end relation extraction from text data using natural language processing, discovering clinical concepts within the texts and their relationships. Current solutions often resolve the problem in two steps: identifying the named entities and classifying any relations between them (Giorgi, J., et al., 2019). These two steps can be regarded as two traditional subtasks, namely named entity recognition (NER) and relation classification (RC), which propagate errors from NER to RC without giving feedback from RC to NER (Bethard, S., et al., 2015, June; Lee, H. J., et al., 2016, June).

Due to the limitation of error propagation between subtasks, we developed the clinical XLNet model based on natural language processing to achieve end-to-end relation extraction from clinical text data.

The structure of the paper has been structured as follows: Section 2 presents the recent works of literature; section 3 depicts the detailed description of the proposed methodology; section 4 deliberates the implementation results; finally, section 5 concludes the paper.

2. Literature Survey

This section provides a survey of some existing research related to clinical relations extraction from electronic health records.

Mahendran, D., & McInnes, B. T. (2021) use relation extraction techniques to examine the relationship between drugs and associated attributes. An outline of three approaches is provided: a rule-based approach, a deep learning approach, and a contextualized language model approach. Experimental results demonstrated that the contextualized language model-based approach outperformed other approaches and reached the highest ADE extraction performance.

Hasan, F., Roy, A., & Pan, S. (2020, November) investigate whether word and sentence embeddings can be used to improve the accuracy of relation extractions using traditional NLP features. In order to extract clinical relationships, different neural network architectures have been explored for combining text embeddings (e.g., Word2Vec and BERT) and traditionally syntactic and semantic features. Comparison between models employing static word embedding (e.g., Word2Vec) and models employing contextual embedding (e.g., BERT). Even though conventional contextual embedding with BERT is very effective, models that combine the technique with traditional syntactic and semantic features perform worse than those that combine conventional contextual embedding with static Word2Vec embedding.

Perera, N., Dehmer, M., & Emmert-Streib, F. (2020) describe approaches to Named Entity Recognition (NER) and Relation Detection (RD), which can be used to determine relationships between proteins and drugs or genes and diseases. A specific biomedical or clinical problem can be summarized using large-scale details integrated into networks, allowing easy data management and analysis.

Liao, W., & Veeramachaneni, S. (2009, June) presented a semisupervised learning algorithm that used CRFs in the NER. Their algorithm provided high-precision labels for unlabeled data based on independent evidence. This allows for the automatic extraction of high-accuracy data without redundant information. A more accurate classifier would be the result of the next iteration.

Jiang, S., Zhao, S., Hou, K., Liu, Y., & Zhang, L. (2019, October) modeled Chinese electronic medical records using a BERT-BiLSTM-CRF model. A pre-trained BERT language model improves word semantics, followed by a biLSTM network with a CRF layer, with word vectors used as inputs. In contrast to previous approaches that aimed to utilize feature engineering and domain knowledge, the BERT model improves semantic representation. At the same time, the BiLSTM network solves the issue of previous methods that relied on feature engineering and domain knowledge. The CRF model is a new approach that focuses on context annotation information instead of other approaches.

Shi, X., et al. (2019) proposed a novel joint deep learning method to recognize clinical entities or attributes and extract entity-attribute relations simultaneously. This method combines two state-of-the-art methods for named entity recognition and relationship extraction, such as bidirectional long short-term memory with conditional random fields and long short-term memory.

Xu, J., He, H., Sun, X., Ren, X., & Li, S. (2018) introduced a unified model which allows semi-supervised learning to learn in-domain unlabeled data by self-training. Large amounts of unlabeled data are combined to enhance the performance of the NER.

Li, F., Zhang, M., Fu, G., & Ji, D. (2017), using a joint neural model to extract biomedical entities and relationships, investigated how knowledge could be extracted from biomedical data. This methodology combines several state-of-the-art neural models for entity recognition and relation classification in natural language processing and text mining. Besides extracting adverse events related to drug-induced diseases, their model also extracted residences related to bacteria.

Throughout the analysis, the use of relation extraction techniques to examine the relationship between drugs and associated attributes requires a lot of manual work

(Mahendran, D., & McInnes, B. T., 2021). The word and sentence embeddings are used to improve the accuracy of relation extractions using traditional NLP features [9]. The clinical problem can be summarized using large-scale details integrated into networks (Perera, N., Dehmer, M., & Emmert-Streib, F., 2020). The high-precision labels for unlabeled data for automatic data extraction (Liao, W., & Veeramachaneni, S., 2009, June). Word semantics, a pre-trained BERT language model, is used, followed by a biLSTM network with a CRF layer, which provokes discrepancy issues (Jiang, S., Zhao, S., Hou, K., Liu, Y., & Zhang, L., 2019, October). A novel joint deep learning method to recognize clinical entities or attributes (Shi, X., et al., 2019). Large amounts of unlabeled data are combined to enhance the performance of the NER, but more memory is required for training (Xu, J., He, H., Sun, X., Ren, X., & Li, S., 2018). A joint neural model to extract biomedical entities and relationships, but there are high possibilities of error propagation (Li, F., Zhang, M., Fu, G., & Ji, D., 2017). A novel relation extraction technique is proposed, elaborated in the next section, to eliminate the issues mentioned above.

3. End-to-End Clinical Relation Extraction

A new framework for assessing patient histories and conducting clinical research has evolved due to the expanding number of clinical reports. Clinical entity recognition and relation extraction are two of the most critical challenges for extracting valuable information from clinical text data. The existing research developed a pipeline for performing entity recognition and relation extraction tasks separately, which could propagate the error from the former task to the latter one, provoking error propagation and performance degradation is provoked. A novel strategy named end-to-end clinical relation extraction has been proposed to overwhelm the issues mentioned above. To deal with the issue of error propagation, both entity recognition and relation extraction

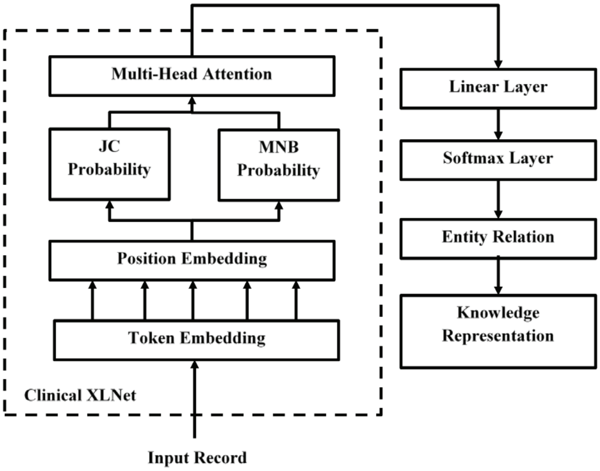


Fig. 1. Block diagram of the proposed end-to-end clinical knowledge discovery strategy

are performed solely rather than separately. Clinical XLNet, a large bidirectional transformer model, has been incorporated as a base model to resolve BERT-based models' discrepancies. Here, the position embedding has been incorporated upon clinical XLNet to take advantage of the entities' positional data of the entities by entity markers with the existing [CLS] vector. To leverage the dependent event relation association and the independent events, the multinomial Naïve Bayes probability function has been performed parallel to the joint conditional probability function. Whereas the multinomial Naïve Bayes leverages the prior probability with conditional probability to determine the probability of the dependent event relation association; furthermore, if the entity pairs are presented in consecutive sentences, the relations between them have been learned through incorporating multi-head attention layer at the top of the clinical XLNet. The transformer outputs are concatenated through a linear layer. Finally, the softmax classifier can be employed to present the entity relation, which might be drug-adverse drug event, drug-reason, etc. At last, the absolute and abundant knowledge discovered from the entity relation extraction will be portrayed in a graphical or statistical form. The block diagram of the proposed methodology is depicted in figure 1.

3.1. Clinical Entity Recognition

The clinical image or clinical text has been utilized for clinical knowledge discovery, whereas we have extracted the knowledge by adopting clinical text data. Initially, the clinical text data has been pre-processed to retain the eminence of the input corpus for data miners to gain knowledge. The pre-processing steps include data cleaning, tokenization, parts of speech tagging, parsing, stemming, and lemmatization. After the data cleaning process, the sentences are split into several tokens, and then the parts of the speech process have been carried out. In order to resolve the structural ambiguity, parsing is carried out, which identifies the constituents.

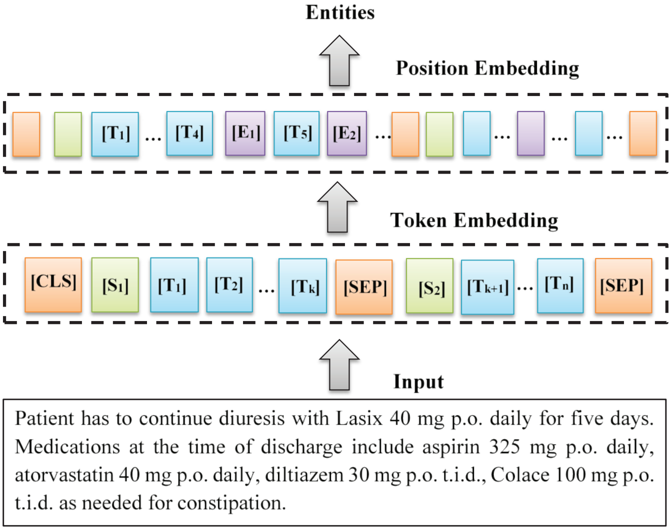
Consequently, stemming and lemmatization have been performed. The procedure for stemming is to remove and replace word suffixes to arrive at a common root for the term. A lemma is a canonical form of a word, whereas a stem may or may not be an actual word. This reduces a word's inflectional forms and certain derivationally related forms to a single base form.

The clinical named entity recognition is the task that is crucial for relation extraction. If the entities are not recognized appropriately, the error of entity recognition was carried forward into the relation extraction task on the knowledge discovery pipeline. This provokes error propagation, implying performance degradation and the erroneous discovery of knowledge. To prevent this, the clinical entity recognition and the relation extraction tasks were performed solely on the single model. Most existing approaches rely on sequence labeling models like BERT with Bi-LSTM or CRF, which suffered from discrepancy issues while tuning with contextual embedding. In order to overcome this, the XLNet, a large bidirectional transformer model, has been adopted as a base model in our research.

The XLNet, on the other hand, has been fine-tuned by utilizing smaller clinical data to maximize the task-specific knowledge to make XLNet a clinical XLNet, where the named entity recognition has been taken place. The embedding layers are mainly contributed to clinical entity recognition. The XLNET embedding should be pre-trained with the clinical notes datasets using Permutation Language Modeling. The permuted language model (PLM) maintains the benefits of autoregressive modeling while simultaneously allowing systems to include bidirectional context. If a sentence is given as $y = (y_1, y_2, y_3 \dots\dots\dots y_m)$with the length of m, there is m! Possible permutations. Symbolize $H_m$as the permutations of set {1,2,3,……m}. For a permutation $h \in H_m$, represents $h_T$as the T-th element in h and $h < T$ as the first T-1 element in h. PLM pre-trains the model by achieving the preceding criteria,

$$logP(y;\theta) = \mathrm{E}_h \in H_m \sum_{T=c+1}^{m} logP(y_{hT}|y_{h<T};\theta) \qquad (1)$$

During the autoregressive pre-training, each predicted token in PLM can only view its previous tokens in a permuted phrase and has no knowledge of the whole sentence's position. The permutation Language modeling has been done in the embedding layer to generate a better clinical embedding from clinical notes. The embedding layer processes are depicted in figure 2. Initially, the pre-processed data has been inputted to the clinical XLNet model, whereas the token embedding layer generates the inputs as tokens for further processing. In clinical XLNet, the [CLS] token has been employed as a classification token; the [SEP] token is for the separation of sentences. However, the [CLS] token utilizes only partial knowledge from the input sequence, which affects the positioning of entities. Thus to resolve this, the position embedding has been incorporated upon clinical XLNet to leverage the position information of the entities by entity markers with the existing [CLS] vector, which contributed to carrying forward the entire knowledge up to the end of the task. In figure 2, the entity tokens are represented as entity markers{$E_1, E_2, \dots E_N$}, which has been appropriately identifying the clinical entities in the input clinical text data.


Entities
Position Embedding
[T1] … [T4] [E1] [T5] [E2] … … … …
Token Embedding
[CLS] [S1] [T1] [T2] … [Tk] [SEP] [S2] [Tk+1] … [Tn] [SEP]
Input

Patient has to continue diuresis with Lasix 40 mg p.o. daily for five days. Medications at the time of discharge include aspirin 325 mg p.o. daily, atorvastatin 40 mg p.o. daily, diltiazem 30 mg p.o. t.i.d., Colace 100 mg p.o. t.i.d. as needed for constipation.

Fig. 2. Embedding Process

As with the entity markers, each entity is preserved until the end of the knowledge discovery process. The clinical entities are drug, class, dosage, frequency, duration, route, and condition. Thus the relation between the drug with the different entities has been extracted by the relation extraction process, which is explained in the next section.

4. Result and Discussion

This section provides a detailed explanation of the implementation results and the evaluation metrics of the proposed system. A comparison has also been provided to demonstrate that the proposed model's performance is superior to existing models.

4.1. Experimental Setup

This work has been implemented in the working platform of python with the following system specification, and the simulation results are discussed below.

Platform	: Python
OS	: Windows 10
Processor	: 64-bit Intel processor
RAM	: 8 GB RAM
Dataset	: N2C2 dataset

4.1.1. Dataset Description

The corpus utilized in the proposed methodology is National NLP Clinical Challenges (n2c2) 2018 dataset. The n2c2 corpus contains the electronic health record of several patients in 10 classes. The medical history, diagnosis, drug name, class of the drug, treatment plans, vaccination dates, allergies, the frequency of the intake of drugs, duration, route, condition, radiological pictures, and laboratory and test results of a patient are all kept in the record. The discharge summaries of patients presented in the corpus are considered for knowledge discovery. The n2c2 corpus contains entity annotations: drug, strength, form, dosage, frequency, route, duration, reason, and ADR. There are eight different kinds of clinical entity relations: strength–drug (severity), form–drug (form), dosage–drug (do), frequency–drug (fr), route–drug (route), duration–- drug (du), reason–drug (reason), ADR–Drug (adverse).

The n2c2 corpus has been employed as the input to the proposed clinical XLNet model for knowledge discovery. The corpus is grouped into 10 classes, each containing several patients' health records that have been initially pre-processed for cleaning, then represented statistically in figure 3.

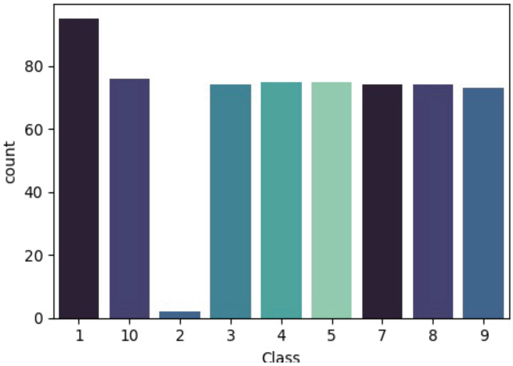
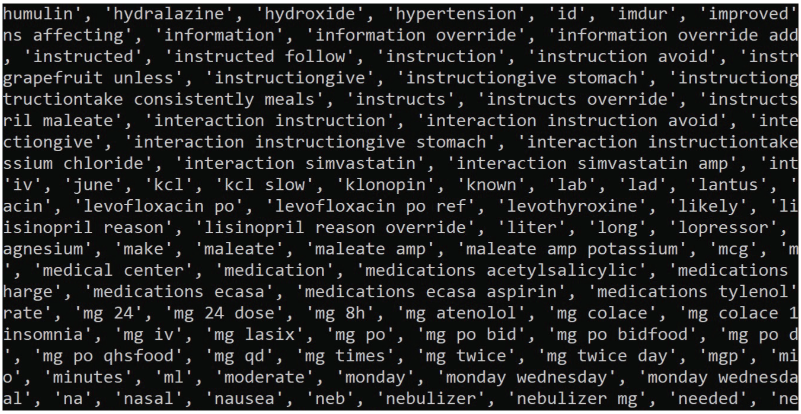
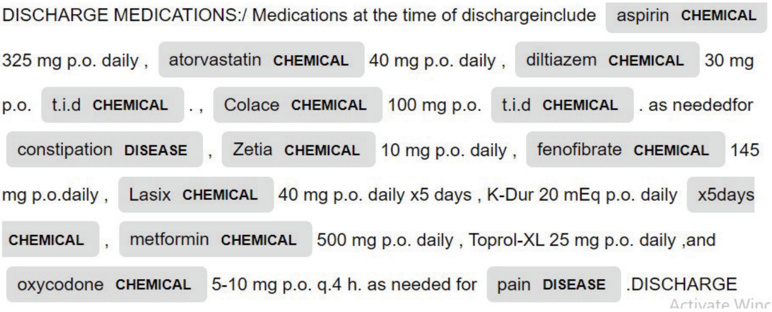


Fig. 3. Statistical representation of n2c2 corpus



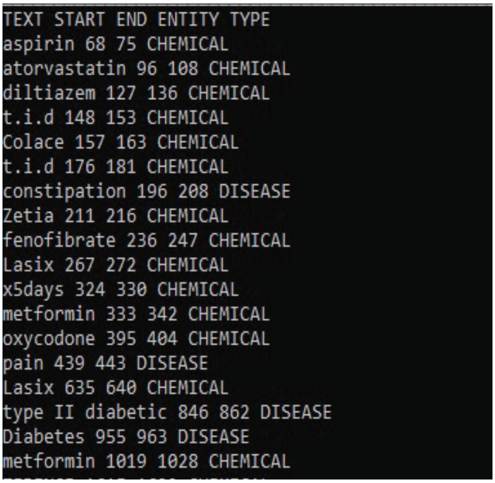
```
humulin', 'hydralazine', 'hydroxide', 'hypertension', 'id', 'imdur', 'improved'
ns affecting', 'information', 'information override', 'information override add
, 'instructed', 'instructed follow', 'instruction', 'instruction avoid', 'instr
grapefruit unless', 'instructiongive', 'instructiongive stomach', 'instructiong
tructiontake consistently meals', 'instructs', 'instructs override', 'instructs
ril maleate', 'interaction instruction', 'interaction instruction avoid', 'inte
ctiongive', 'interaction instructiongive stomach', 'interaction instructiontake
ssium chloride', 'interaction simvastatin', 'interaction simvastatin amp', 'int
'iv', 'june', 'kcl', 'kcl slow', 'klonopin', 'known', 'lab', 'lad', 'lantus', '
acin', 'levofloxacin po', 'levofloxacin po ref', 'levothyroxine', 'likely', 'li
isinopril reason', 'lisinopril reason override', 'liter', 'long', 'lopressor', 
agnesium', 'make', 'maleate', 'maleate amp', 'maleate amp potassium', 'mcg', 'm
', 'medical center', 'medication', 'medications acetylsalicylic', 'medications 
harge', 'medications ecasa', 'medications ecasa aspirin', 'medications tylenol'
rate', 'mg 24', 'mg 24 dose', 'mg 8h', 'mg atenolol', 'mg colace', 'mg colace 1
insomnia', 'mg iv', 'mg lasix', 'mg po', 'mg po bid', 'mg po bidfood', 'mg po d
', 'mg po qhsfood', 'mg qd', 'mg times', 'mg twice', 'mg twice day', 'mgp', 'mi
o', 'minutes', 'ml', 'moderate', 'monday', 'monday wednesday', 'monday wednesda
al', 'na', 'nasal', 'nausea', 'neb', 'nebulizer', 'nebulizer mg', 'needed', 'ne
```

Fig. 4. Tokenization



DISCHARGE MEDICATIONS:/ Medications at the time of dischargeinclude aspirin **CHEMICAL**
325 mg p.o. daily , atorvastatin **CHEMICAL** 40 mg p.o. daily , diltiazem **CHEMICAL** 30 mg
p.o. t.i.d **CHEMICAL** . , Colace **CHEMICAL** 100 mg p.o. t.i.d **CHEMICAL** . as neededfor
constipation **DISEASE** , Zetia **CHEMICAL** 10 mg p.o. daily , fenofibrate **CHEMICAL** 145
mg p.o.daily , Lasix **CHEMICAL** 40 mg p.o. daily x5 days , K-Dur 20 mEq p.o. daily x5days
CHEMICAL , metformin **CHEMICAL** 500 mg p.o. daily , Toprol-XL 25 mg p.o. daily ,and
oxycodone **CHEMICAL** 5-10 mg p.o. q.4 h. as needed for pain **DISEASE** .DISCHARGE

Fig. 5 (a)



```
TEXT START END ENTITY TYPE
aspirin 68 75 CHEMICAL
atorvastatin 96 108 CHEMICAL
diltiazem 127 136 CHEMICAL
t.i.d 148 153 CHEMICAL
Colace 157 163 CHEMICAL
t.i.d 176 181 CHEMICAL
constipation 196 208 DISEASE
Zetia 211 216 CHEMICAL
fenofibrate 236 247 CHEMICAL
Lasix 267 272 CHEMICAL
x5days 324 330 CHEMICAL
metformin 333 342 CHEMICAL
oxycodone 395 404 CHEMICAL
pain 439 443 DISEASE
Lasix 635 640 CHEMICAL
type II diabetic 846 862 DISEASE
Diabetes 955 963 DISEASE
metformin 1019 1028 CHEMICAL
```

fig. 5 (b)

Fig. 5. Entity Relation Extraction

For knowledge discovery, the clinical entity recognition task is the one that has to be focused on to enhance the proposed model's efficiency. To do that, the tokenization takes place with token embedding, which is depicted in figure 4. On the obtained tokens, the incorporation of position embedding has been identifying the position of the entities and the entity relation with the drug, which is portrayed in figure 5. In figure 5(a), the entity types marked in the input record have been illustrated. In figure 5(b), the identified entity relations with the drug are listed in the output window with the position of entities, whereas the entity drug is indicated as chemical.

5. Conclusion

In this paper, a novel end-to-end clinical relation extraction strategy has been proposed with clinical XLNet. The proposed model has considered the extraction of entity pairs in consecutive sentences, and this provides vast knowledge from the clinical text data, which has been implemented with the n2c2 corpus.

References

Bethard, S., et al. (2015, June). Semeval-2015 task 6: Clinical tempeval. In *proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (pp. 806-814).

Bose, P., et al. (2021). A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts. *Applied Sciences, 11*(18), 8319.

Friedman, C., et al. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association, 1*(2), 161-174.

Giorgi, J., et al. (2019). End-to-end named entity recognition and relation extraction using pre-trained language models. *arXiv preprint arXiv:1912.13415.*

Hasan, F., Roy, A., & Pan, S. (2020, November). Integrating Text Embedding with Traditional NLP Features for Clinical Relation Extraction. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 418-425). IEEE.

Jiang, S., Zhao, S., Hou, K., Liu, Y., & Zhang, L. (2019, October). A BERT-BiLSTM-CRF model for Chinese electronic medical records named entity recognition. In *2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA)* (pp. 166-169). IEEE.

Lee, H. J., et al. (2016, June). UTHealth at SemEval-2016 task 12: an end-to-end system for temporal information extraction from clinical notes. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 1292-1297).

Li, F., Zhang, M., Fu, G., & Ji, D. (2017). A neural joint model for entity and relation extraction from biomedical text. *BMC bioinformatics, 18*(1), 1-11.

Liao, W., & Veeramachaneni, S. (2009, June). A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing* (pp. 58-65).

Mahendran, D., & McInnes, B. T. (2021). Extracting Adverse Drug Events from Clinical Notes. In *AMIA Annual Symposium Proceedings* (Vol. 2021, p. 420). American Medical Informatics Association.

Perera, N., Dehmer, M., & Emmert-Streib, F. (2020). Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental* *biology,* 673.

Shi, X., et al. (2019). Extracting entities with attributes in clinical text via joint deep learning. *Journal of the American Medical Informatics Association, 26*(12), 1584-1591.

Tang, B., Cao, H., Wu, Y., Jiang, M., & Xu, H. (2013, April). Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. In *BMC medical informatics and decision making* (Vol. 13, No. 1, pp. 1-10). BioMed Central.

Xu, J., He, H., Sun, X., Ren, X., & Li, S. (2018). Cross-domain and semisupervised named entity recognition in chinese social media: A unified model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26*(11), 2142-2152.

Zhang, R., Zhao, P., Guo, W., Wang, R., & Lu, W. (2022). Medical named entity recognition based on dilated convolutional neural network. *Cognitive Robotics, 2,* 13-20.

Submitted: 03.09.2021
Accepted: 25.11.2021