International Journal of Engineering & Technology, 7 (4.5) (2018) 297-302

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Evaluation of Named Entity Recognition Algorithms Using Clinical Text Data

J. Manimaran[1]* and T. Velmurugan[2]

[1]*Research Scholar, Research & Development Centre, Bharathiar University, Coimbatore, India*
[2]*Associate Professor, PG and Research Dept. of Computer science, D. G. Vaishnav College, Chennai, India*
**Corresponding author E-mail thavasimaniraj@gmail.com[1], velmurugan_dgvc@yahoo.co.in[2]*

Abstract

Named Entity Recognition (NER) is one of the most important research areas in the field of medical. Presently, most of the clinical NER research is based on two approaches as Knowledge Engineering (KE) and Machine Learning (ML). KE is used a word lookup table approach and ML is known as supervised learning approach. The aim of this work is to evaluate a recent algorithm in KE and ML approaches using various clinical text databases. Therefore, the NOBLE Coder and Clinical Named Entity Recognition (CliNER) algorithms are selected, NOBLE Coder is depended on KE approach and CliNER is ML approach. The two algorithms will be described and compared its performance on three openly available datasets that is obtained from Medical Information Mart for Intensive Care II (MIMIC II), Pittsburgh Medical Center, and i2b2 2010 challenge. Among these datasets, the annotated data are included which is used to detect the highest sensitivity and specificity on each algorithm. The randomly distributed patient reports were taken as input data to these algorithms. By executing these algorithms, the information is extracted and which classified into predefined concept types, for example medical problems, treatments and tests. The accuracy of both algorithms is calculated using standard measures. The taken two algorithms are analyzed based on the produced results. Finally, the best among two is suggested for better use in clinical data.

Keywords: *Natural Language Processing; Text Mining; Information Extraction; Medical Text Data.*

1. Introduction

Natural Language Processing (NLP) is mainly used in a variety of clinical research to extract useful information from unstructured texts (e.g., pathology reports, discharge summaries). The well-known clinical NLP applications are information extraction (IE), information retrieval, text generation, user interfaces, and machine translation [1]. In these applications, IE is the process of extracting structured information from unstructured and/or semi-structured machine-readable documents. In medical field, IE is a process to identify medical terms and phrases from unstructured text documents. For example, the noun phrase of a disease is highlighted in the following statement “The patient is a 28-year-old woman who is **HIV positive** for two years”. In many IE research, biomedical named entity recognition technique is widely considered to be one of the most important steps in the following works searching the diagnostic code of diseases [2], adverse drug reactions extraction [3], abbreviations extraction [4], de-identification [5], obesity [6], medication [7], relation extraction [8], coreference resolution [9], temporal relations extraction [10], etc. Named Entity Recognition (NER) is also called as concept extraction/recognition. The term ‘NER’ has been focused primarily on two challengeable tasks: 1) identification of clinical concepts and 2) classification of these concepts into predefined entity types (e.g. medical problem, tests, and treatments). Clinical NER uses two main approaches, the symbolic and statistical approach. The symbolic is a KE process which can be used in morphological, lexical, syntactic, semantic, pragmatic, and discourse. In contrast, statistical is based on a ML approach. The objective of this work is to evaluate both the KE and ML approach by using various clinical text databases. In this case, the NOBLE Coder and Clinical Named Entity Recognition (CliNER) algorithms are selected for KE and ML approach. Three openly available datasets are being used for this research as test set which are taken from Medical Information Mart for Intensive Care II (MIMIC II), Pittsburgh Medical Center, and Informatics for Integrating Biology to the Bedside (i2b2) 2010 challenge. The preprocessing task is not important to use these datasets due to the algorithms can apply directly in the clinical text documents. But, the post-processing is important to use in the algorithm results. The gold standard test is an evaluation process that is used to define the patient’s true disease state. It supports the development of all NLP algorithms. Here, the selected datasets includes annotated data for gold standard test, which is generated by domain experts. This data is not same in all datasets and so the preprocessing task is proposed in this work to set the data in a common format.

Finally, the performance of both algorithms would be calculated using the extrinsic metrics “BCubed”. This evaluation metric is based on comparisons between the output of our NLP systems and human annotated data. The remainder of this paper is structured as follows: In section 2 presents the applications of clinical NER. The materials and methods for the proposed work are described in section 3. The experimental and evaluation results are showed in section 4. Section 5 discusses about the algorithm performances and finally, section 6 concludes this research work.

2. Applications of Clinical NER

In NLP research, NER is primarily focused on medical domain due to more than 50% of researches have already made significant progress on clinical NER. The major challenges of clinical NER are diagnostic code detection, adverse drug reactions (ADRs) extraction, abbreviations extraction, de-identification, obesity, medication, relation extraction, coreference resolution, temporal relations extraction, protein name recognition etc. This section shows the clinical NER applications on both KE and ML approach. Ira et al. were analysed three different methods for diagnostic code detection in the radiology reports [2]. The ML based diagnostic code detection methods have studied and developed evaluation metrics by Adler perotte et al. [12]. In ADRs extraction, Heidemann et al. developed a novel text searching tool to capture idiosyncratic drug-induced liver illness cases from electronic medical record system. This was based on KE approach [3]. Combination of deep recurrent neural networks and conditional random fields were used to develop a new model for ADRs extraction [13]. The biomedical abbreviations are extracted through the link-topic model algorithm (statistical model) [4]. In addition, Yonghui et al. were organized open-source framework for clinical abbreviation recognition and disambiguation (including ML, clustering, and KE) [14]. De-identification is the process of removing private health information from medical discharge records. This same method evaluated using the experimental dataset “i2b2 2006 challenge” [5]. Obesity is now present in the most cases. Classifying obesity and its comorbidities is called the obesity challenge [6]. In order to this challenge is addressing by both the KE and ML approaches. The challenge was targeted that identification of medications, their dosages, modes (routes) of administration, frequencies, durations, and reasons for administration in discharge summaries.

Some authors have proposed that the state-of-the-art techniques for medication challenge [7, 15]. In 2010, i2b2/VA workshop distributed the annotated corpora and primarily focused on the relation classification task in clinical texts [8]. In 2011, the i2b2/VA was shared a resource for coreference challenge, it is most important to determine whether two concepts are coreference for example the coreference chain are highlighted in the sentence “She was scheduled to receive a temporal artery biopsy, but she never followed up on that testing” [9]. In addition the temporal relations are another challenge to identify time-related information in the medical free texts. Evaluating temporal relations in clinical text [10] have also done by Weiyi Sun et al. Kaoru Yamamoto et al. find the issues of biomedical NER in protein name recognition (likewise tokenization ambiguity, changing nomenclature, feature engineering) and they also proposed a morphological analyser to support biomedical text processing.

3. Materials and Methods

Traditionally the IE task has been used to detect the boundaries and identify the categories of clinical entities, and map them to concepts in standardized terminologies. This section is to describe the details of proposed evaluation by using the following materials and methods: the selected datasets, pre-processing, medical NER algorithms, post processing and evaluation metric. Overall process is illustrated in figure 1. In this figure, there are three components are important: the first is the NER module, in which the specified algorithms apply and detect entities from clinical text. The second is the post processing module, which determines correct entities based on the annotated data. The third is the evaluation module, which calculates the performance of each algorithm using the standard measures.

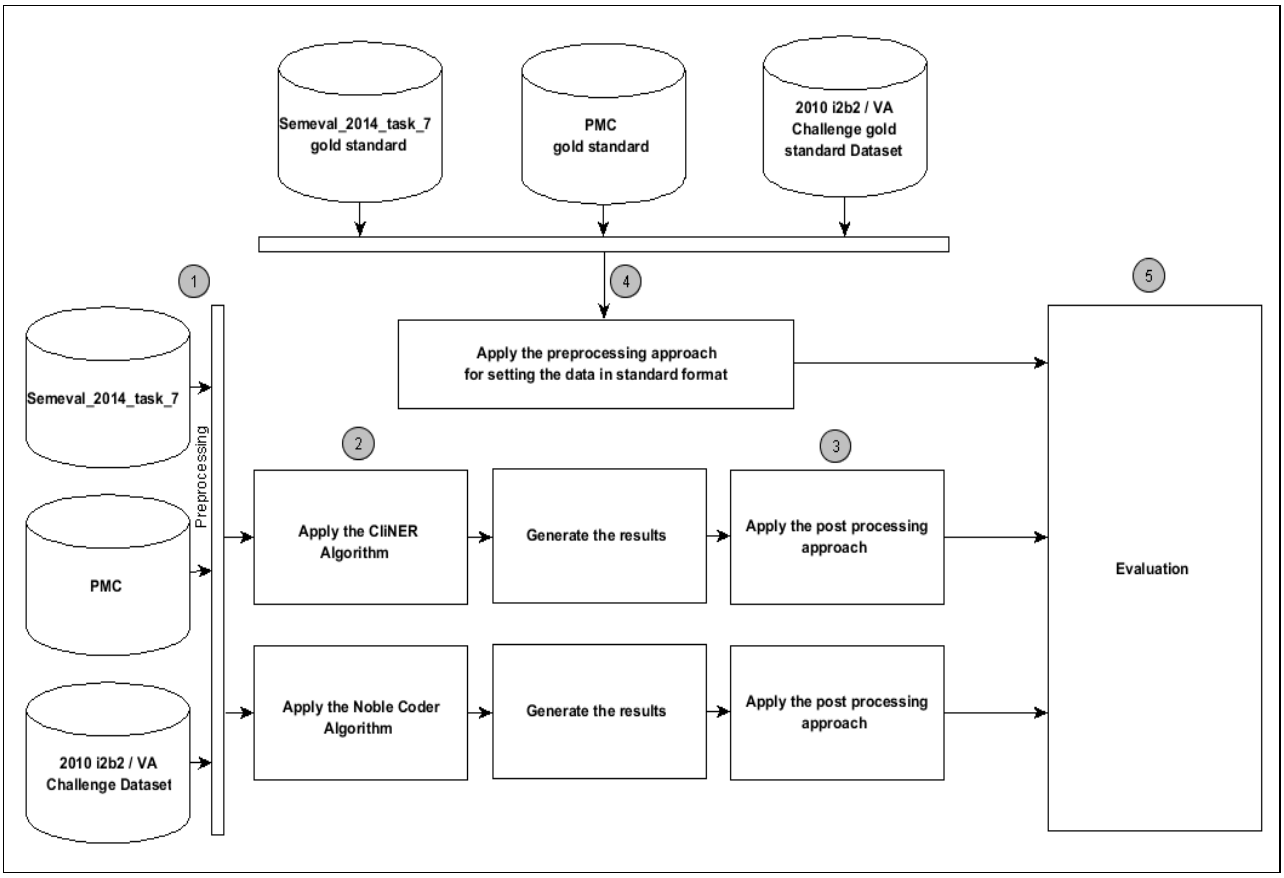


Fig. 1: Architecture of proposed method

3.1. Dataset Description

Today, the datasets along with gold standard data is existed and shared only a few numbers for biomedical NER. This research is aimed to access openly available dataset in which MIMIC II, Pittsburgh Medical Center, and i2b2 2010 challenge. In MIMIC II dataset, trail data on semantic evaluation 2014 task 7 is selected for disorder mentions and normalized to an UMLS Concepts. The second database is used in this study include discharge summaries between January 1, 1993 and December 31, 1995 at two medical ICU’s at the University of Pittsburgh Medical Center. In addition, the third dataset was obtained from i2b2 2010 challenge by the data use agreement. It was classified into two: train and test set. Table 1 show that the numbers of clinical documents are going to be used in this research. The gold standard data have in different formats and it needs the pre-processing task to set the data in a common format. Therefore, the data pre-processing is done in the annotated data.

Table 1: Description of Data set

Datasets	Documents
Semeval_2014_task_7	4
Pittsburgh Medical Center	2376
i2b2 2010 beth training data	73
i2b2 2010 partners training data	97
i2b2 2010 test data	256

3.2. Methods of Pre-Processing

The selected NER algorithms can effectively deal with an unstructured clinical text document; hence it does not require any external pre-processing steps. But, the problem is if the dataset had already in structured data, then that dataset must pre-process and send it to the chosen NER algorithms. In this research, the second dataset is structured format and the dataset is manually pre-processed before starting the NER algorithm.

3.3. Medical NER Algorithms

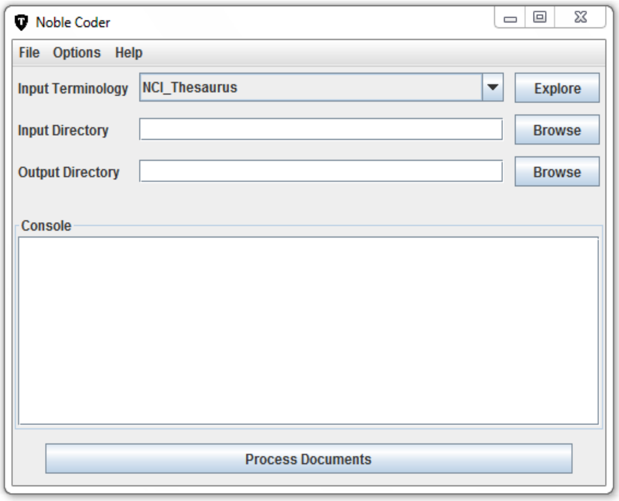
Medical NER algorithms, namely NOBLE Coder and CliNER are applied in this research. The characteristics of these algorithms will be described in the following subsection.

NOBLE Coder algorithm: NOBLE Coder is a term-to-concept matching algorithm which is implemented in Java using two hash tables: a word-to-terms (WT) table and a term-to-concepts (TC) table. This is an open source and easily integrated into the existing system UIMA (unstructured information management architecture) or GATE (general architecture for text engineering). This algorithm involves two major tasks,

1) Terminology building process
2) Concept recognition process

Terminology building process is basically designed to minimize the user effort when construct an input terminologies. Figure 2 shows the screen shot of NOBLE coder application. For example, NOBLE Coder uses a bundled terminology importer user interface so users can easily import custom terminologies in multiple formats (RRF, OWL, OBO, and BioPortal). In WT table, words are normalized using an approach that is similar to the method used by SPECIALIST NLP tools. Each normalized term is then mapped to its corresponding concept in the TC table. In parallel, each word from a given term is mapped to a set of normalized terms that contain it in the WT table. To perform the concept recognition process, input text is broken into a set of normalized words and stop words are excluded. The word set is then ranked by frequency of associated terms. Each word is looked up in the WT table to find terms that are associated with the word and include all of the other words in the input text. This term is then added to a candidate list. Once all of the words in the input text have been processed and a set of candidate terms generated, each candidate term is looked up in the TC table and its concept is added to the results queue [16].

CliNER algorithm: Clinical Named Entity Recognition (CliNER) is a machine learning based algorithm for the extraction of named entities from clinical text. CliNER is also an open-source NLP application for clinical NER and it's a two-pass supervised ML system. The first pass identifies concept boundaries using linear chain conditional random fields. The second pass assigns clinical concept types to the phrases identified in the first pass, where is using the classification approach support vector machines [17].



Input Terminology NCI_Thesaurus

Fig. 2: Screen shot of NOBLE coder application

Concept boundary detection uses general text features such as word, stem, and part-of-speech n-grams and includes that domain specific feature as GENIA, UMLS metathesaurus and semantic network. In concept type identification, annotated data with desired clinical concepts can be used to build a model which can then be used to recognize similar concepts in raw text. This state-of-the-art algorithm is extensible and easy-to-use architecture. This is implemented in Python using the following packages sklearn, CRFsuite, and LibSVM. In addition, table 2 shows the common difference between two algorithms.

Table 2: Difference between Noble Coder and CliNER algorithm

Noble Coder	CliNER
Not Extract the concept start and end information	Extracting the concept start and end information
Dictionary Based	Non-dictionary based
More than three types are used for concepts	Three types are used for concepts
It includes an intractive terminology builder tool	It is not include that any terminology building process
Training is not required and therefore it does not need any training data	Training is required and therefore it needs the training data
It is an user interface application	It is a command line based application
It need not to split a sentence before submitting input data	It needs to split a sentence before submitting input data

3.4. Post-Processing

Post processing is one of the essential steps to remove an irrelevant data from the algorithm results. Output of each algorithm should be compared with one another by using the structure query language on Microsoft SQL Server. For each algorithm, the results are post processed and imported it into the Microsoft SQL Server. Then, using the SQL joins, the imported data can be compared with the gold standard data.

3.5. Pre-Processing the Gold Standard Data

In the above said datasets, the gold standard data's are not in a same format. Table 3 shows that the gold standard database and its annotation format. Caption is the number of parameters used in every datasets. First dataset, the preprocessing task is performed by hand due to the dataset contain less input documents. The second dataset is a less manual work compare than previous dataset, where it was already arranged the concepts and their sentences in a single row.

Table 3: Annotated file format

Databases	Delimiter	Captions
Semeval_2014_task_7	||	File name, disease_disorder, disease_code, start, end
PMC	Tab	Serial number, concepts, sentences, assertions
Beth training data	||	Concept text, concept type, assertion value
Partners training data	||	Concept text, concept type, assertion value
i2b2 2010 test data	||	Concept text, concept type, assertion value

Finally, the manual preprocessing is a lot of time consuming tasks in the third dataset; therefore an automatic preprocessing function is developed in this work to set the data in a single row. After the preprocessing, the extracted concepts of gold standard data can be shown in the table 4.

Table 4: Preprocessed gold standard data

Datasets	Gold Standard
Semeval_2014_task_7	234
Pittsburgh Medical Center	2376
i2b2 2010 beth training data	10296
i2b2 2010 partners training data	6229
i2b2 2010 test data	31161

3.6. Evaluation Metric

Evaluation metric is normally used for finding the algorithms efficiency. It is classified into two subjects: intrinsic and extrinsic. Intrinsic metric can be mostly applied to determine whether a concept is closely related to one class to another class. On the other hand, extrinsic metric is based on the comparisons between the computer and human generated concepts (gold standard). This research chooses an extrinsic measure that proposed by Enrique et al. They were actually proposed the new metric BCubed for text clustering algorithms [18]. We used a simple extended BCubed measure implemented in python, which include precision, recall and F-score.

4. Results

In most of the computer research, there have been developed a numerous medical information technologies those have yielded suboptimal results and meet user dissatisfaction when implemented in practice. The result of this study is divided into two: before and after post processing. The actual results (i.e. before post processing) of two algorithms are summarized in table 5. There are exposed much unwanted data while comparing with the gold standard data. See that the results of Noble Coder algorithm have higher than the CliNER algorithm and the reason behind that it was retrieved all concepts, including more number of semantic types for example diagnostic procedure, disease or syndrome, etc. Even the Noble coder can be used more types for the medical concepts extraction, this study is specifically focused within a semantic type name as "Finding". On before post processing, the percentage of CliNER algorithm is 80%, 1%, 80%, 32%, 65% and Noble coder is 387%, 7%, 484%, 245%, 405%.

Table 5: Results before post processing

Datasets	CliNER	Noble Coder
Semeval_2014_task_7	282	1549
Pittsburgh Medical Center	2411	15768
i2b2 2010 beth training data	5847	35315
i2b2 2010 partners training data	3147	23791
i2b2 2010 test data	16759	103608

Figure 3 illustrates the results of two NER algorithms as early as the post processing is started. After post processing, the results of two algorithms are shown in table 6. It was removed an irrelevant data when compared with the gold standard data. Note that the results of CliNER algorithm increased among the all i2b2 datasets while compare than the Noble coder algorithm.

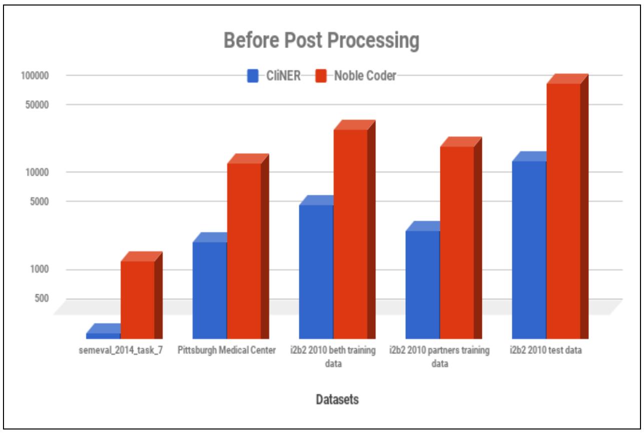


Fig. 3: Results of before post processing

After post processing, the percentage of CliNER algorithm is 6%, 0.78%, 80%, 22%, 68% and Noble coder is 46%, 2%, 9%, 4%, 9%.

Table 6: Results after post processing

Datasets	CliNER	Noble Coder
Semeval_2014_task_7	22	182
Pittsburgh Medical Center	1849	4901
i2b2 2010 beth training data	5808	677
i2b2 2010 partners training data	2153	432
i2b2 2010 test data	17508	2222

Figure 4 shows graphically the results of taken algorithms as shown in table 6. In the post processing task, the percentage of unwanted data extracted by CliNER is 12%, 0.30%, 0.01%, 0.46%, 0% and Noble coder is 8%, 2%, 51%, 54%, 46%.

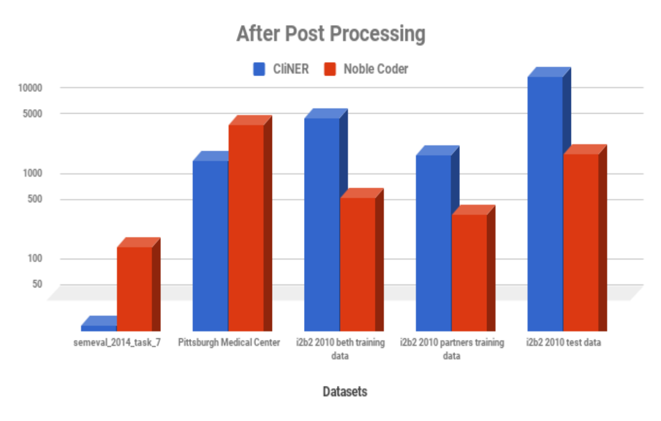


Fig. 4: The algorithm results on after post processing

5. Discussion

The evaluation task is important to identify the best clinical NER algorithm. After the step 3 and 4 (see at figure 1), the evaluation task performs using the advanced statistical measures, which is described in table 7. It is interesting to note that the CliNER algorithm has better than the Noble Coder algorithm, which is showed in the Bcubed results precision, recall, and F-Score. The TP (true positive) results indicate that the extracted information is correct and FP (false positives) results indicate that the extracted information is incorrect. In addition, error analysis is also important to notify that the percentage of error occurred into the two algorithms. Table 8 illustrates the errors using the classified data as retrieved, relevant, irrelevant, excess information, not retrieved. Finally the error percentage has been calculated by using the formula,

$$\text{Error \%} = \text{Irrelevant} / \text{Retrieved} \qquad (1)$$

At the table, excess data is if the algorithm have retrieved more information than the gold standard data (excess = retrieved - gold standard). Missing data's could be highlighted within the column as "not retrieved" and these values were derived from manual data minus relevant data. A symbol '-' indicate that the excess result is how much data have missed between the retrieved and manual results? and not retrieved data is how much data has additionally retrieved between the manual and relevant data?. The classification of excess and not retrieved data can be useful to decide whether the retrieved results are sufficient to the gold standard data. Figure 5 summarizes the results in table 8. As shown in the graph 5, the number of errors is increased in CliNER algorithm during the input dataset "Semeval_2014_task7". It is interesting to note that the errors of CliNER algorithm have decreased when using the other three datasets. Along with it is important to note that the Noble Coder algorithm could be suffered in a large number of errors. The reason is this algorithm was directly worked with the test data; whereas the CliNER algorithm should be used first the training data (to build up a model) and second the test data (to validate the model built). Although the CliNER algorithm has been better than the Noble Coder algorithm, since which did not prove its reliability (to reduce the errors) as we expected. A recent study of NER systems using the CRAFT and ShARe corpus showed that Noble Coder achieved acceptable F- Score when compared to various knowledge engineering systems [16].

Table 7: Evaluation results

Datasets	Algorithms	TP	FP	Bcubed		
				Precision	Recall	F-score
Semeval_2014_task_7	CliNER	22	260	1	0.03	0.05
	NC	182	1367	1	0.18	0.3
PMC	CliNER	1849	562	1	0.1	0.19
	NC	4901	10867	1	0.06	0.11
Beth training data	CliNER	5808	39	1	0.3	0.46
	NC	677	34638	1	0.07	0.12
Partners training data	CliNER	2153	994	1	0.17	0.29
	NC	432	23359	1	0.15	0.27
i2b2 2010 test data	CliNER	8872	7887	1	0.17	0.3
	NC	2222	101386	1	0.09	0.17

Table 8: Errors of two algorithms

DB	Algorithms	Retrieved	Relevant	Irrelevant	Excess	Not Retrieved	Error %
1	CliNER	282	22	260	48	212	0.92
	NC	1549	182	1367	1315	52	0.88
2	CliNER	2411	1849	562	35	527	0.23
	NC	15768	4901	10867	13392	(-)2525	0.68
3	CliNER	5847	5808	39	(-)4449	4488	0.006
	NC	35315	677	34638	25019	9619	0.98
4	CliNER	3147	2153	994	(-)3082	4076	0.31
	NC	23791	432	23359	17562	5797	0.98
5	CliNER	16759	8872	7887	(-)14402	22289	0.47
	NC	103608	2222	101386	72447	28939	0.97

This study did not specifically compare KE systems, whereas our results provide another type of NER system based on ML and which tested against five different datasets. The findings of this research significantly differ from other comparative studies [19-20], our results do not support to confirm their observation, in fact it compared different algorithm in different dataset which had mainly aimed to find that the best algorithm between KE and ML approach.

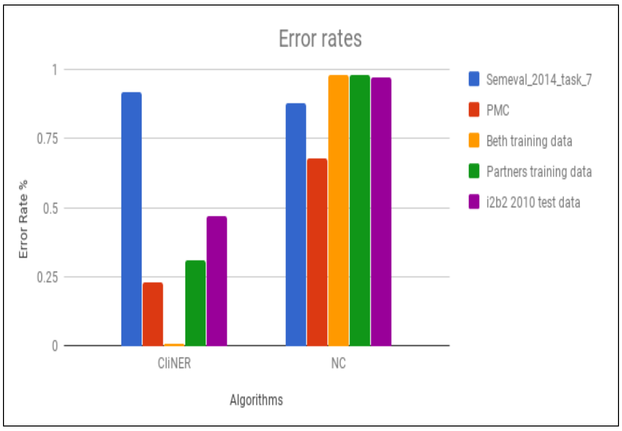


Fig. 5: Error rate of two algorithms

Additionally, no previous study has evaluated for these two algorithms. Although there were some advantages into this, the experiments prove that both algorithms have failed reducing the erroneous data. This study was unsuccessful in proving that which concepts have been extracted from which sentences. For example, if we extract characters from string, we should know where to start the extraction and where to end the extraction. This research may have three limitations. The first is limited algorithms and datasets were used for this evaluation. The second is the post processing step as we used is not automated. The third is CliNER algorithm is only running on Linux operating system.

6. Conclusion

This work has evaluated clinical NER algorithms for KE and ML approach. This work used two algorithms namely Noble coder (KE) and CliNER (ML), which have been successfully tested against five databases. In addition, this work highlights that the importance of post processing methods. After post processing, those results were evaluated by the extrinsic measures. The results of this study indicate that the CliNER algorithm significantly better than Noble coder algorithm based on the performance of the algorithms. The present study was limited in two ways. First, it presents only one algorithm for KE and ML approach. Second, the extracted data is verified by the annotation file, but they did not check more deeply sentence by sentence. The reason behind that the Noble coder is not writing which concepts have been extracted from which sentences. However, the CliNER algorithm would not be best at before post processing. Finally, this research suggests that the presented ML based algorithm is useful in further examination. Moreover, semantic analysis is also very important and will be considered in future research.

Acknowledgement

We would like to thank Mr. A. Dharmarajan, Dept. of Computer Science, Ayya Nadar Janaki Ammal College, Sivakasi, India, Mr. S. Shahul Hameed and Mr. A. Purushothaman from Sree Balaji Medical College and Hospital, Chennai, India.

References

[1] Edward H. Shortliffe and James J. Cimino, “Biomedical Informatics: Computer Applications in Health Care and Biomedicine (Health Informatics)”, Springer-Verlag New York, Inc., 2006.

[2] Ira Goldstein, Anna Arzumtsyan, and Ozlem Uzuner, “Three Approaches to Automatic Assignment of ICD-9-CM Codes to Radiology Reports”, AMIA Annual Symposium Proceedings, pp. 279–283 (2007).

[3] Lauren Heidemann, James Law, and Robert J. Fontana, "A text searching tool to identify patients with idiosyncratic drug-induced liver injury", Digestive diseases and sciences, Vol. 62, No. 3, pp. 615-625, 2017.

[4] Seonho Kim and Juntae Yoon, "Link-topic model for biomedical abbreviation disambiguation", Journal of biomedical informatics, Vol. 53, pp. 367-380, 2015.

[5] Ozlem Uzuner, Yuan Luo, and Peter Szolovits, "Evaluating the state-of-the-art in automatic de-identification", Journal of the American Medical Informatics Association, Vol. 14, Issue 5, pp. 550-563, 2007.

[6] Ozlem Uzuner, "Recognizing obesity and comorbidities in sparse data", Journal of the American Medical Informatics Association, Vol. 16, Issue 4, pp. 561-570, 2009.

[7] Ozlem Uzuner, Imre Solti, and Eithon Cadag, "Extracting medication information from clinical text", Journal of the American Medical Informatics Association, Vol. 17, Issue 5, pp. 514-518, 2010.

[8] Ozlem Uzuner et al., "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text", Journal of the American Medical Informatics Association, Vol. 18, Issue 5, pp. 552-556, 2011.

[9] Ozlem Uzuner et al., “Evaluating the state of the art in coreference resolution for electronic medical records”, Journal of the American Medical Informatics Association, Vol. 19, Issue 5, pp. 786-791, 2012.

[10] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner, "Annotating temporal information in clinical narratives", Journal of biomedical informatics, Vol. 46, pp. S5-S12, 2013.

[11] Kaoru Yamamoto et al., "Use of morphological analysis in protein name recognition", Journal of Biomedical Informatics, Vol. 37, Issue 6, pp. 471-482, 2004.

[12] Adler Perotte et al., “Diagnosis code assignment: models and evaluation metrics”, Journal of the American Medical Informatics Association, Vol. 21, Issue 2, pp. 231–237, 2014, https://doi.org/10.1136/amiajnl-2013-002159

[13] Elena Tutubalina and Sergey Nikolenko, "Combination of Deep Recurrent Neural Networks and Conditional Random Fields for Extracting Adverse Drug Reactions from User Reviews", Journal of Healthcare Engineering, 2017.

[14] Yonghui Wu et al., "A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD)", Journal of the American Medical Informatics Association, Vol. 24, Issue e1, pp. e79-e86, 2017.

[15] Jon Patrick and Min Li, "High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge", Journal of the American Medical Informatics Association, Vol. 17, Issue 5, pp. 524-527, 2010.

[16] Eugene Tseytlin et al., "NOBLE–Flexible concept recognition for large-scale biomedical natural language processing", BMC bioinformatics, Vol. 17, Issue 32, 2016.

[17] William Boag et al., "CliNER: A lightweight tool for clinical named entity recognition", AMIA Joint Summits on Clinical Research Informatics (poster), 2015.

[18] Enrique Amigo et al., "A comparison of extrinsic clustering evaluation metrics based on formal constraints", Information retrieval, Vol. 12, Issue 4, pp. 461-486, 2009.

[19] Neil Ireson et al. "Evaluating machine learning for information extraction", Proceedings of the 22nd international conference on Machine learning. ACM, pp. 345-352, 2005.

[20] N. Kanya, T. Ravi, and S. Geetha, "A comparative study of Information Extraction tools used for Biological database", Sustainable Energy and Intelligent Systems (SEISCON 2011), International Conference, pp. 886-890, 2011.