2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Comparison of ACM and CLAMP for Entity Extraction in Clinical Notes

Fatemeh Shah-Mohammadi, Wanting Cui, Joseph Finkelstein, *Member, IEEE*

***Abstract*— Rapid increase in adoption of electronic health records in health care institutions has motivated the use of entity extraction tools to extract meaningful information from clinical notes with unstructured and narrative style. This paper investigates the performance of two such tools in automatic entity extraction. In specific, this work focuses on automatic medication extraction performance of Amazon Comprehend Medical (ACM) and Clinical Language Annotation, Modeling and Processing (CLAMP) toolkit using 2014 i2b2 NLP challenge dataset and its annotated medical entities. Recall, precision and F-score are used to evaluate the performance of the tools.**

***Clinical Relevance*— Majority of data in electronic health records (EHRs) are in the form of free text that features a gold mine of patient's information. While computerized applications in healthcare institutions as well as clinical research leverage structured data. As a result, information hidden in clinical free texts needs to be extracted and formatted as a structured data. This paper evaluates the performance of ACM and CLAMP in automatic entity extraction. The evaluation results show that CLAMP achieves an F-score of 91%, in comparison to an 87% F-score by ACM.**

I. Introduction

In Electronic Health Records (EHR) or Electronic Medical Records (EMR), patients' information are recorded in either a structured format (e.g., diagnosis codes, medications and laboratory results) or an unstructured format (e.g., clinical text notes in the form of discharge summaries, radiology notes and progress notes). Clinical text notes contain vast amounts of information about the patient such as detailed patient conditions and prescribed medications. Due to its unstructured nature, the information from clinical notes needs to be extracted and then categorized for further utilization and analysis in daily healthcare settings and research [1]. One solution is to employ domain experts to manually perform the information extraction. However, this solution can be time-consuming and error-prone [2]. As a result, automated systems that can extract information with high accuracy and efficiency are necessitated. Recently, natural language processing (NLP) has been widely used to realize such automated systems. This domain and its subdomain, information extraction (IE), aims to automatically extract information from unstructured data [3].

Named entity recognition (NER) is a subtask within the field of IE that deals with recognition of entities in a free text. Another subtask associated with IE is category classification and relationship extraction (RE). Category classification focuses on classifying the extracted entities into predefined categories such as person names, medication and test treatment procedure. While RE focuses on identifying the relation amongst extracted entities [4,5]. Two types of methods mainly employed in these subtasks are the rule-based method and machine learning. The former is the predominant approach applied to the clinical texts. This method is a collection of handcrafted rules that requires collaboration with domain experts [1]. However, the machine learning method produces better results as long as a large dataset is available to train the machine learning model. There are also some systems named as "hybrid systems" that utilize both methods [6,7].

Nowadays, various IE systems have been developed to extract information from clinical notes [8]. Amazon Comprehend Medical (ACM) is one of such systems that has been recently developed by Amazon Web Services (AWS). ACM's machinery is driven and powered by state of the art deep learning models, and is trained and updated in line with the evolvement of end-user requirements. It provides multiple access modes such as console and software development kits supporting various programming languages and platforms. ACM also links detected entities to standardized medical knowledge bases such as Rx-Norm and ICD10-CM via ontology linking operations.

Clinical language annotation, modeling and processing (CLAMP) toolkit is also an NLP-based clinical IE system that enables automatic extraction and encoding of entities in clinical notes. It is not only a high performance NLP system but also an interactive development environment (IDE) for building customized clinical NLP solutions.

As mentioned, recognition and extraction of medical entities such as diseases, medications and treatments plays a critical role for patients and medical research. Extracted medical entities also form the basis for other tasks such as disease correlation, classification and diagnosis [9-11]. Due to the significance of medical entity extraction, this paper aims to compare the entity extraction performance of CLAMP and ACM. For this project, we worked with the 2014 i2b2 NLP challenge dataset for identifying heart disease and its risk factors in diabetic patients [12]. The automated extraction resulted from CLAMP and ACM was evaluated against the expert's annotations.

The rest of this paper is organized as follows: Section II

F. Shah-Mohammadi is with Icahn School of Medicine at Mount Sinai, 1770 Madison Ave, 2nd Fl, New York, NY, 10035 USA (phone: 5856228990, e-mail: fatemeh.shah-mohammadi@mountsinai.org).

W. Cui is with Icahn School of Medicine at Mount Sinai, 1770 Madison Ave, 2nd Fl, New York, NY, 10035 USA (e-mail: wanting.cui@mssm.edu).

J. Finkelstein is with Icahn School of Medicine at Mount Sinai, 1770 Madison Ave, 2nd Fl, New York, NY, 10035 USA (e-mail: Joseph.Finkelstein@mssm.edu).

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

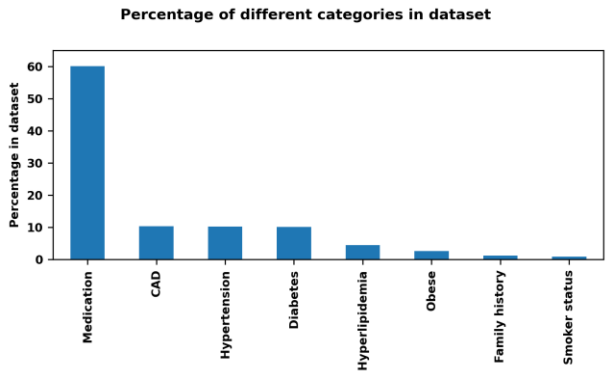

Figure 1. Statistics of categories in dataset.

focuses on describing the corpus used in this paper and presenting our data cleaning pipeline developed to separate the actual narrative text from annotations. This section also describes the evaluation metrics based on which the performance of ACM and CLAMP are compared. The results of comparison are presented in Section III, followed by discussion and conclusion in Sections IV and V, respectively.

II. METHODS

Amazon Comprehend Medical was released by Amazon Web Service (AWS) in 2018 to automatically extract clinical concepts from clinical notes. ACM leverages a deep learning-based system which constitutes two Long Short Term Memory (LSTM) encoders at the character and word level and a single tag decoder. Transfer learning has been further added to this base framework to overcome the constraint of limited access to medical data for training purposes and to enable generalizability of the model across different medical specialties [13-15]. ACM can detect entities in the following categories:

- Anatomy: this tag relates to the body parts, systems and their location.
- Medical condition: this tag involves diagnosis of medical condition and the corresponding acuity, signs and symptoms.
- Medication: this tag includes medication and its associated dosage, frequency, duration, and form for the patient.
- Test treatment procedure: this tag deals with the diagnosis testing and treatment procedures used to determine a medical condition.
- Protected health information: this tag relates to the patient's personal information.

ACM also links those entities tagged as medication to normalized concept identifiers from the RxNorm database from the US national library of medicine (NLM).

Despite the fact that several general-purpose IE system such as ACM have been developed to process the clinical texts, studies have shown that end users need to take substantial effort to adopt existing NLP systems [16]. Moreover, users often witnessed reduced performance when an existing general-purpose IE system is applied without customization beyond its original purpose (e.g., when different types of clinical notes are fed to the system) [17]. CLAMP is a newly developed NLP system that provides end users with a graphic user interface to help them build their own customized NLP pipelines for their individual applications that require substantial NLP skills. In particular, CLAMP presents a pipeline-based architecture that builds NLP systems from multiple components [18]. In this work we considered CLAMP's default clinical pipeline. This pipeline decomposes into multiple components as follows: tokenizer, part-of-speech (POS) tagger, section identifier, named entity recognizer, assertion classifier, attribute recognizer, concept mapper, temporal recognizer, and temporal relation (more details regarding functionality of each component can be found in [18] and references therein). Similar to ACM, this pipeline can also link the extracted entities to normalized concept identifiers from the RxNorm databases.

A. Dataset

We consider track2 i2b2 2014 NLP challenge dataset. This dataset consists of 521 medical texts distributed in XML documents and annotated by the experts. Each XML formatted document was composed of the actual narrative text and the annotations. First, we separated the narrative text from the annotations. It should be noted that annotations contain the original text recognized and tagged by the expert, its character offset and its associated category. There are 8 unique categories listed as: "*diabetics and coronary artery disease (CAD)*" showing the presence and progression of disease, "*hypertension*", "*hyperlipidemia*", "*obesity*", "*medication*", "*smoking status*", and "*family history*". We then imported the XML formatted annotations into a relational database to facilitate data analytics. In particular, we used Python's built-in *xml.etree* module to parse the XML data into a tree with different nodes. We also defined a function that iterates over the nodes and extracts the tags and attributes associated with each node.

Both IE systems considered in this paper can link their extracted entities to normalized concept identifiers from the RxNorm database. Since RxNorm provides normalized names and unique identifiers only for medicines and drugs, amongst 8 categories mentioned above we only considered entities categorized as "*medication"*. Moreover, according to Fig. 1 that shows the percentage of entities associated with each category, entities tagged as "*medication*" account for around 60% of the annotations.

Table.1 shows the list of entities considered in this paper. The first column of this table lists the original text within the document that recognized and annotated by the experts. The second column aggregates the identified Rx-Norm codes for each entity. It should be noted that CLAMP outputs not only the RXNorm code but also the generic code, while ACM delivers only the RXNorm code. As a result, we have only listed in this table the detected RXNorm codes by the tools. The last column lists preferred name associated to each RXNorm code.

TABLE I. LIST OF ENTITIES AND THEIR LINKED RXNORM CODES

Annotated Entities	RxNorm Code	Preferred Name
Atenolol	[1202], [197381]	Atenolol, Atenolol 50 MG Oral Table
Norvasc	[58927]	Norvasc
Lipitor	[153165]	Lipitor
Aspirin	[1191], [317300]	Aspirin
Metoprolol	[6918]	Metoprolol
Glucophage	[151827]	Glucophage
Toprol	[865575]	Toprol
Lisinopril	[29046]	Lisinopril
Pravachol	[203333]	Pravachol
Zocor	[196503]	Zocor
Nifedipine	[7417]	Nifedipine
Zestril	[196472]	Zestril
Lovastatin	[6472]	Lovastatin
Pravastatin	[42463], [904481]	Pravastatin, Pravastatin Sodium 80 MG Oral Tablet
Isosorbide	[6057]	Isosorbide
Labetalol	[6185]	Labetalol
Zebeta	[221002]	Zebeta
Coreg	[216221], [686926]	Coreg, Carvedilol 3.125 MG Oral Tablet [Coreg]
Accupril	[72210]	Accupril
Glucotrol	[203680]	Glucotrol

B. Evaluation Metrics

The experts' annotations have been considered as a gold standard to evaluate the automatic entity extraction of the ACM and CLAMP. We scored entity recognition performance of the tools based on not only the *text*, whether the character offsets of the extracted entity and the original text annotated by human match exactly, but also based on whether the associated category is correct [19]. Within the dataset, 350 unique entities have been categorized as medication, among which only 123 were categorized the same by not only ACM but also CLAMP. Out of 123 entities, we randomly considered 20 entities for which both tools provided the same RxNorm code. We have used recall (or sensitivity), precision and F-score measured metrics to evaluate the results [20]. We calculate recall, precision and F-score for each entity type and then their macro-average measures are provided. The programming language for all analysis was Python 3.8.

III. RESULTS

The results have been shown in Table 2. According to this table, the averages for the recall, precision and F-score with CLAMP were 0.88, 0.90, and 0.91, respectively. With ACM, the averages for the same measures were 0.86, 0.94, and 0.87, respectively. In comparison with ACM, CLAMP showed better performance by around 2% for the average recall and 4.6% for the average F-score. On the other hand, in comparison to CLAMP, ACM achieved higher average precision by 4.5%. Both tools scored the lowest recall values for "Accupril" (0.33) meaning that for 3 occurrences of this entity in dataset they were able to detect only 1. ACM showed a low recall value for "Pravachol" (0.39) too, while CLAMP perfectly extracted the entire number of occurrences identified by the annotators for the same entity (i.e. recall equals 1). The next lowest recall for both tools is associated to "Toprol" with the value of 0.5. This recall value means that only half of the occurrences of this entity in dataset were identified by the tools.

Both tools perfectly identified the entire number of occurrences for the first and second most frequent entities, i.e. "Lisinopril" and "Atenolol", in the dataset. For the third most frequent entity, "Aspirin", recall for ACM is 1 while it is 0.99 for CLAMP.

Regarding the least frequent entity "Zebeta", the recall value for ACM and CLAMP is 1, while for the next least frequent entity, i.e. "Accupril", and for the same measure both tools achieved low value of 0.33. The tools achieved much higher results for the third least frequent entity "Lovastatin" with a perfect recall value of 1.

IV. DISCUSSION

Considering the results shown in Table 2, in comparison with ACM, CLAMP had better performance in terms of recall and F-score with 2% and 4.6% higher values, respectively. Amongst the three least frequent entities in Table 2, both tools were able to perfectly identify two of them. For three most frequent entities, CLAMP resulted in an average recall value of around 0.997, while the same metric measure for ACM is 1. It means that ACM performs better in identifying the most frequent entities. Since in this study we only considered the category "*medication*", in the future studies we will evaluate performance of these tools in extracting entities belonged to the other remaining categories.

V. CONCLUSION

Majority of data in EHR are in the form of free text notes which feature a gold mine of information. The information from these notes must be extracted and categorized to be later utilized for clinical decision support, quality improvement and research. Therefore, an automated system will be necessary in order to parse medical information with high

TABLE II. SUMMARY OF EVALUATION

Entities annotated by experts	Frequency of occurrences (sample size equals 1251)	CLAMP			ACM		
		Recall	Precision	F_score	Recall	Precision	F_score
Atenolol	211	1	0.91	0.95	1	0.93	0.96
Norvasc	60	0.80	1	0.89	1	0.90	0.95
Lipitor	185	1	0.99	0.99	1	0.84	0.91
Aspirin	195	0.99	1	0.99	1	0.94	0.97
Metoprolol	69	0.72	1	0.84	0.67	1	0.80
Glucophage	60	0.85	1	0.92	1	1	1
Toprol	36	**0.50**	1	0.67	**0.50**	1	0.67
Lisinopril	225	1	0.89	0.94	1	0.86	0.92
Pravachol	23	1	0.92	0.96	**0.39**	1	0.56
Zocor	34	0.82	1	0.9	1	0.83	0.91
Nifedipine	23	0.91	1	0.95	0.83	1	0.91
Zestril	53	0.96	1	0.98	1	0.81	0.89
Lovastatin	4	1	1	1	1	1	1
Pravastatin	34	0.82	1	0.90	1	0.92	0.96
Isosorbide	7	1	0.88	0.94	1	0.88	0.94
Labetolol	8	1	**0.80**	0.89	1	**0.80**	0.89
Zebeta	2	1	1	1	1	1	1
Coreg	7	0.86	1	0.92	0.86	1	0.92
Accupril	3	**0.33**	1	0.50	**0.33**	1	0.50
Glucotrol	12	1	1	1	0.67	1	0.80
Average		0.88	0.90	0.91	0.86	0.94	0.87

The lowest values for *recall and precision are in bold*

efficiency and accuracy. In this paper, we compared the automatic entity extraction performance of two IE systems: CLAMP and ACM. The result of our conducted experiment on 20 entities showed that CLAMP outperforms ACM by 2% for the average recall and 4.6% for the average F-score. While ACM showed better performance in terms of average precision by achieving 4.5% higher score in comparison with CLAMP. For three most frequent entities, Clamp resulted in an average recall value of around 0.997, while ACM achieved 1. Since a good IE system is the one that can correctly catch as higher number of entities as possible and CLAMP showed better performance in terms of the average recall, we will proceed with CLAMP for real-world applications.

REFERENCES

[1] Wang, Yanshan, et al., "Clinical information extraction applications: a literature review," *Journal of biomedical informatics 77,* pp. 34-49, 2014.

[2] Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C. and Chute, C.G., "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, 17(5), pp.507-513,2010.

[3] Small, Sharon Gower, and Larry Medsker, "Review of information extraction technologies and applications," in *Neural computing and applications* 25, no.3, pp: 533-548, 2014.

[4] Mikheev, Andrei, Marc Moens, and Claire Grover, "Named entity recognition without gazetteers," in *Ninth Conference of the European Chapter of the Association for Computational Linguistics*. 1999.

[5] Bach, Nguyen, and Sameer Badaskar, "A review of relation extraction," in *Literature review for Language and Statistics II 2*, pp: 1-15, 2007.

[6] Doan, Son, et al., "Recognition of medication information from discharge summaries using ensembles of classifiers," in *BMC medical informatics and decision making* 12.1, pp: 1-10, 2012.

[7] Tang, Buzhou, et al., "A hybrid system for temporal information extraction from clinical text," in *Journal of the American Medical Informatics Association 20*, no. 5, pp: 828-835, 2013.

[8] Reátegui R, Ratté S. , "Comparison of MetaMap and cTAKES for entity extraction in clinical notes," *BMC medical informatics and decision making.* 2018;18(3):13-9.

[9] Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, Brunak S, "Using electronic patient records to discover disease correlations and stratify patient cohorts," in *PLoS Comput Biol*. 2011;7(8):1–10.

[10] Yildirim P, Çeken Ç, Hassanpour R, Tolun MR, "Prediction of similarities among rheumatic diseases," in *JMed Syst.* ,36(3), pp:1485–90, 2012.

[11] Bejan CA, Xia F, Vanderwende L, Wurfel MM, Yetisgen-Yildiz M, "Pneumonia identification using statistical feature selection," in *JAMIA.* , 19(5), pp:817–23,2012.

[12] Stubbs, Amber et al., "Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2," in *Journal of biomedical informatics.* 58 Suppl(Suppl):S67-77, 2015.

[13] P. Bhatia, B. Celikkaya, M. Khalilia and S. Senthivel, "Comprehend Medical: A Named Entity Recognition and Relationship Extraction Web Service," *18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Boca Raton, FL, USA, pp. 1844-1851, 2019.

[14] Jin, M. et al., "Improving hospital mortality prediction with medical named entities and multimodal learning," *arXiv preprint arXiv:1811.12276*.

[15] Bhatia, P. et al., "Dynamic Transfer Learning for Named Entity Recognition," *arXiv preprint arXiv:1812.05288*.

[16] Uzuner, Ö., Gururaj, A.E., Bayer, S., Aberdeen, J., Rumshisky, A. and Pakhomov, S., "Ease of adoption of clinical natural language processing software: an evaluation of five systems," *Journal of biomedical informatics*, 58, pp:189-196, 2015.

[17] Chapman, W.W., Nadkarni, P.M., Hirschman, L., D'avolio, L.W., Savova, G.K. and Uzuner, O., "Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions," 2011.

[18] Soysal, Ergin, et al., "CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines," *Journal of the American Medical Informatics Association 25.3*, pp: 331-336, 2018.

[19] Yadav, V. and Bethard, S., "A survey on recent advances in named entity recognition from deep learning models," *arXiv preprint arXiv:1910.11470*, 2019.

[20] Chinchor, N. and Sundheim, B.M., "1993. MUC-5 evaluation metrics," *In Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore*, Maryland, August 25-27, 1993.