Clinical Entity Extraction: Comparison between MetaMap, cTAKES, CLAMP and Amazon Comprehend Medical

Lu Bai
School of Computing
Ulster University
Belfast, UK
l.bai@ulster.ac.uk

Zhibao Wang
School of Computer and Information Technology
Northeast Petroleum University
Daqing, China
wangzhibao@nepu.edu.cn

Maurice D. Mulvenna
School of Computing
Ulster University
Belfast, UK
md.mulvenna@ulster.ac.uk

Raymond Bond
School of Computing
Ulster University
Belfast, UK
rb.bond@ulster.ac.uk

***Abstract*— In recent years, electronic health records (EHRs) have been adopted widely and there is an increasing need to extract useful clinical information from free-text clinical notes. In this work, we compare the performance of the clinical entity extraction tools including MetaMap, cTAKES, CLAMP and Amazon Comprehend Medical. The clinical notes dataset we use in this work is i2b2 Obesity Challenge dataset. The experiments are designed to extract a list of the clinical entities related to obesity symptoms or clinical conditions using four different clinical entity extraction tools. The medical entities were manually annotated by two obesity experts in the dataset which are used as the ground truth. The evaluation has been done by using evaluation metrics including precision, recall, and F1-score and comparison has been made of different clinical entity extraction tools and APIs. The results show that MetaMap has the highest recall (0.61) and F1-score (0.70) and CLAMP has the highest precision (0.98) of the averages for all the selected clinical conditions. However, for certain clinical conditions, cTAKES and Amazon Comprehend Medical outperform other tools. The results demonstrate that these clinical entity extraction tools are able to automatically and accurately extract useful information from the clinical notes.**

Keywords*— *Amazon Comprehend Medical, CLAMP, cTAKES, Clinical entity extraction, Clinical notes, MetaMap

I. INTRODUCTION

The use of electronic health records (EHRs) has increased significantly in recent years [1]. It is important to obtain information and knowledge from EHRs to support the care and clinical decision making for the secondary use of the EHRs in clinical research [2]. EHRs consist of both structured data such as the diagnostic code, physiological measurements, and medication. EHRs also contain unstructured data which is mostly the free text clinical notes including discharge summary, radiology notes and clinical letters to name but a few [3]. The analysis and visualization of the structured data from EHRs is straight forward. However, the manual process for extracting useful information from unstructured free text clinical notes (e.g. individual clinical terms) are tedious and error-prone and typically requires clinical domain knowledge [4]. Information extraction (IE) in natural language processing (NLP) is widely used to extract the concepts and entities from free-text documents. A number of IE systems or tools are expert-based systems which are able to extract the entities and locations of entities and even relationships between different entities from text [5].

There are a range of IE tools and APIs for clinical entity extraction. MetaMap [6] and cTAKES [7] are two popular IE tools which have been used widely in a lot of applications to extract clinical concepts from free-text documents including identifying concepts from discharge summaries [8], radiology reports [9], and medical social media documents [10]. MetaMap is developed based on the Unified Medical Language System (UMLS) and provides users with access to the biomedical concepts in UMLS Metathesaurus [6]. cTAKES is a clinical IE system which combines the rule-based method and machine learning techniques for clinical narrative processing [7]. Previously, a number of studies have also made the comparison between the MetaMap and cTAKES [3], [11]. CLAMP (Clinical Language Annotation, Modeling, and Processing) [12] is a relatively new clinical NLP toolkit currently available to research for free use. It is built on Apache Unstructured Information Management Architecture™ (UIMA) framework and uses a pipeline-based architecture. CLAMP provides the GUI interface named CLAMP-GUI, which aims to build customized NLP pipelines. Amazon Comprehend Medical [13] is a relatively new service for clinical entity extraction which is released by Amazon in November 2018. Amazon Comprehend Medical service uses machine learning to extract clinical concepts from free-form medical text. The user can quickly extract useful information for medical conditions and medication dosages by using a simple API call to Amazon Comprehend Medical service [13].

Automatically extracting clinical entities is of great importance as useful information can be obtained to gain an immediate insight from a long clinical free-text note [2]. The automatic extraction of clinical concepts are widely used in a range of applications benefiting both clinicians and patients, especially in the area of patient phenotyping [14] and clinical decision support [15]. With emergence of the deep learning based algorithms, a number of studies use the extracted clinical entities rather than the raw clinical notes as the inputs to the deep learning models, which further improves the

 2021 32nd Irish Signals and Systems Conference (ISSC), 2021, pp. 1-6, doi: 10.1109/ISSC52156.2021.9467856.

performance of the model [16]. This work extends a previous work [3] and includes the comparison of more clinical entities extraction tools. In this paper we aim to evaluate a number of clinical entities extraction tools to provide further reference in automatic extraction of clinical entity for other researchers. Using the dataset from i2b2 (Informatics for Integrating Biology to the Bedside) 2008 Obesity challenges, we conduct the experiments on clinical entity extraction using different clinical entities extraction tools. Subsequently, evaluation is made for each of the clinical conditions related to obesity based on the judgement from the obesity experts.

II. METHODS

A. Dataset

In this work, we choose to use i2b2 2008 Obesity dataset [17] to evaluate the performance of four different clinical entity extraction tools. This dataset also contains the ground truth of two obesity experts who were from the Massachusetts General Hospital (MGH) Weight Centre. They had provided textual annotations for each of the clinical documents in the dataset. The annotations from the two obesity experts are used as the ground truth to evaluate different clinical entity extraction tools used in the experiment. In total, 611 clinical notes are used in this work.

B. Unified Medical Language System (UMLS)

UMLS [18] is developed by US National Library of Medicine and it is a repository of biomedical vocabularies. Some of the clinical entity extraction tools e.g. cTAKES, MetaMap and CLAMP require the use of UMLS resources. The Metathesaurus is one of the most important and biggest component of the UMLS and it is the vocabulary database that contains information about biomedical and health-related concepts. In the Metathesaurus, each unique clinical concept is assigned a concept unique identifier (CUI) which is associated with one or more semantic type and a CUI can represent terms that equivalent in meaning [19]. The user will need to register with UMLS and obtain a UMLS license with an API key in order to use these tools.

In this work, we aim to evaluate the clinical concepts extraction performance on 15 clinical conditions which are linked to obesity. Table I shows the 15 clinical conditions we aim to extract from the clinical notes and their corresponding CUIs.

TABLE I. SUMMARY OF THE EXTRACTED CLINICAL CONDITIONS USING DIFFERENT CLINICAL ENTITY EXTRACTION TOOLS

Clinical Conditions	UMLS CUI
Asthma	C0004096
CAD	C1956346
CHF	C0018802
Depression	C0011570
Diabetes	C0011849
Gallstones	C0008320
Gout	C0018099
Hypercholesterolemia	C0020443
Hypertension	C0020538
Hypertriglyceridemia	C0020557
OA	C0029408
Obesity	C0028754
OSA	C0520679
PVD	C0085096
Venous Insufficiency	C0042485

C. Clinical Entity Extraction Tools

In this work, we use four different clinical entity extraction tools to extract medical terms. In the following subsections, we will discuss each of these tools in more details. All of these tools can provide the location of the extracted clinical entities.

1) cTAKES

cTAKEs uses UMLS to extract the clinical entities and is able to provide users with CUI, Extracted Clinical Terms, Type Unique Identifier (TUI), SNOMED CT code [20], negations, categories etc. A named entity to -1 indicates that it has been negated. Table II below shows an example of extracted information by using cTAKES.

2) MetaMap

MetaMap is an NLP tool for recognizing UMLS concepts in text which was developed by National Library of Medicine (NLM). It is a knowledge extraction system which maps the biomedical text to the UMLS Metathesaurus [6], [21]. In this work, MetaMap lite [22] is used to extract the clinical terms, and it is a lighter named-entity recognizer. Table III shows the extracted clinical terms and their corresponding CUIs and categories.

TABLE II. AN EXAMPLE OF EXTRACTED CUI AND CLINICAL TERMS PROVIDED BY CTAKES

CUI	TUI	Extracted Clinical Terms	Category	SNOMED CT code
C0002871	T047	Anemia	Anemia	271737000
C0011849	T047	Diabetes	Diabetes Mellitus	73211009
C0011881	T047	diabetic nephropathy	Diabetic Nephropathy	127013003
C0038454	T047	stroke	Cerebrovascular accident	230690007
C0005823	T040	blood pressure	Blood Pressure	75367002
C0277919	T046	venous stasis	Postthrombotic Syndrome	71897006

TABLE III. AN EXAMPLE OF EXTRACTED CUI AND CLINICAL TERMS PROVIDED BY METAMAP

CUI	Extracted Clinical Terms	Category
C0002871	Anemia	Anemia
C0027627	Diabetes	Diabetes Mellitus
C0011881	diabetic nephropathy	Diabetic Nephropathy
C0038454	stroke	Cerebrovascular accident
C0005823	blood pressure	Blood Pressure
C0277919	venous stasis	Postthrombotic Syndrome

3) CLAMP

CLAMP is a new clinical NLP tool which allows the users to build their own NLP pipelines for different applications. Moreover, there are two versions available including a command-line system named CLAMP-CMD and a GUI based system named CLAMP-GUI. Similarly, as MetaMap and cTAKES, it can output CUI, extracted clinical terms and semantic categories as shown in Table IV.

TABLE IV. AN EXAMPLE OF EXTRACTED CUI AND CLINICAL TERMS PROVIDED BY CLAMP

CUI	Extracted Clinical Terms	Semantic Category
C0002871	Anemia	problem
C0027627	Diabetes	problem
C0011881	diabetic nephropathy	problem
C0038454	stroke	problem
C0005823	blood pressure	test
C0277919	venous stasis	problem

4) Amazon Comprehend Medical

Compared with the above three clinical entity extraction tools, Amazon Comprehend Medical does not provide CUIs. In order to obtain the CUIs of the extracted clinical concepts, we have used the UMLS REST API [23] to return a list of CUIs from the clinical entities obtained using Amazon Comprehend Medical (Table V). Amazon Comprehend Medical also provides a probability of its correctness for each of the extracted clinical terms.

TABLE V. AN EXAMPLE OF EXTRACTED CUI AND CLINICAL TERMS PROVIDED BY AMAZON COMPREHEND MEDICAL

CUI	Extracted Clinical Terms	Sematic Category	Score
C0002871	Anemia	MEDICAL_CONDITION	0.988
C0005823	blood pressure	TEST_TREATMENT_PROCEDURE	0.998

III. EXPERIMENTS

A. Experimental Process

The experimental process is shown in Figure 1 below. Four different clinical entity extraction tools are used to extract the clinical entity on the 611 clinical documents in the i2b2 dataset. Then the extracted terms from all four different clinical entities extraction tools are compared with the ground truth from the judgement of the obesity experts to further evaluate the performance of each tool on different clinical conditions.

B. Evaluation Metrics

In this work, obesity experts' manual annotations are used as the gold standard in evaluating the performance for different clinical entity extraction tools. To evaluate each clinical entity extraction tool, three evaluation metrics are selected for our experiments, including precision, recall, and F1-score. The metrics are defined in the equations as follows:

$$precision = \frac{TP}{TP+FP} \quad (1)$$

$$recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1\ score = 2 * \frac{precision*recall}{precision+recall} \quad (3)$$

Where TP is true positive, FP is false positive and FN is False Negative of the CUIs.

IV. RESULTS

The clinical conditions extraction has been done at a document level. Table VI presents the total number of documents of each clinical conditions extracted from the patients' clinical records using the four different clinical entity extraction tools and the annotations made by two obesity experts. It is noted that cTAKES has extracted the highest number for the total number of clinical conditions. The total number of the annotations made by the experts are higher than all of that extracted clinical entities from the clinical entity extraction tools. It also can be seen from Table VI that for some clinical conditions, e.g. Hypertension, Obesity, OSA, the clinical entity extraction tools extract similar number compared with the ground truth gained from the two experts. However, for some conditions e.g. Diabetes, Venous Insufficiency, there is a significant discrepancy between the annotations from clinical experts and the clinical entity extraction tools.

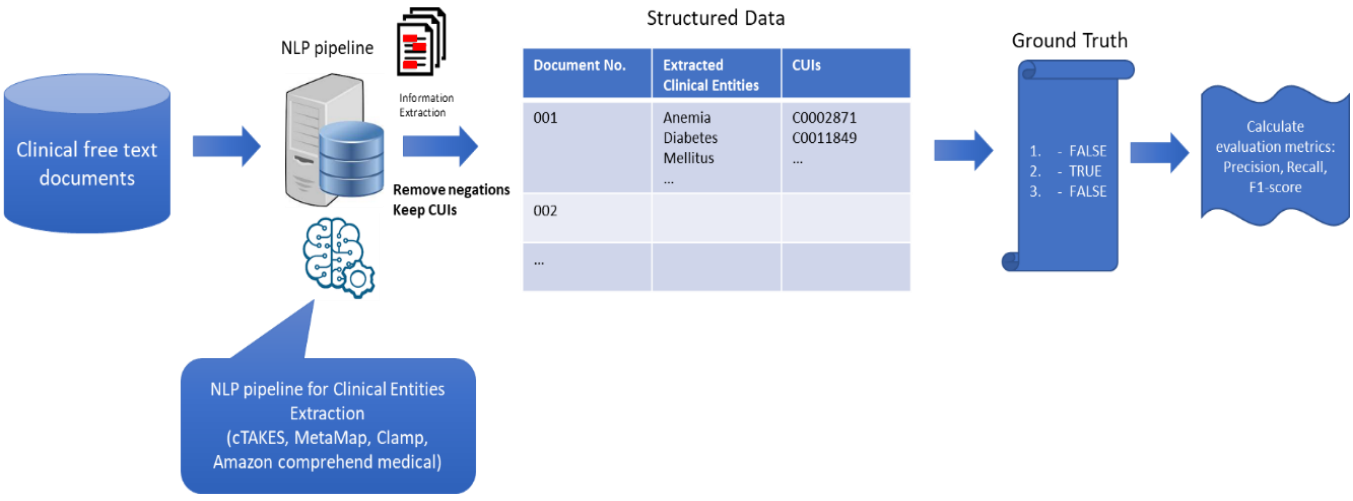

Fig. 1. Clinical entities extraction process

TABLE VI. SUMMARY OF THE EXTRACTED CLINICAL CONDITIONS USING DIFFERENT CLINICAL ENTITY EXTRACTION TOOLS

Clinical Conditions	Ground truth 1	Ground truth 2	MetaMap	cTAKES	CLAMP	Amazon
Asthma	70	75	50	70	62	70
CAD	325	333	234	309	16	1
CHF	239	243	205	139	195	209
Depression	122	86	116	0	83	84
Diabetes	396	399	224	330	214	195
Gallstones	87	91	32	0	56	61
Gout	78	71	69	69	64	67
Hypercholesterolemia	262	246	116	146	134	144
Hypertension	428	442	419	423	409	433
Hypertriglyceridemia	33	15	6	7	6	8
OA	98	89	48	56	52	57
Obesity	239	245	215	227	180	212
OSA	84	88	62	63	64	60
PVD	87	83	71	70	70	69
Venous Insufficiency	44	14	2	1	2	1
Total	2592	2520	1869	1910	1607	1671

[a.] Ground truth 1 and Ground truth 2 are from the annotation of two obesity experts

TABLE VII. SUMMARY OF THE EVALUATION FOR DIFFERENT CLINICAL ENTITY EXTRACTION TOOLS

Clinical Conditions	MetaMap			cTAKES			CLAMP			Amazon		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Asthma	0.92	0.66	0.77	0.93	**0.93**	**0.93**	**0.95**	0.84	0.89	0.94	0.88	0.91
CAD	0.94	0.67	0.78	0.94	**0.9**	**0.92**	**1**	0.05	0.09	0	0	0
CHF	**1**	**0.84**	**0.92**	**1**	0.57	0.73	**1**	0.8	0.89	**1**	0.83	0.91
Depression	0.68	0.65	0.66	0	0	0	**0.89**	0.61	0.72	0.74	**0.71**	**0.73**
Diabetes	0.96	0.54	0.69	**0.97**	**0.81**	**0.88**	0.98	0. 53	0.69	0.96	0.49	0.65
Gallstones	**1**	0.37	0.54	0	0	0	**1**	0.64	0.78	**1**	**0.66**	**0.8**
Gout	0.96	**0.85**	**0.90**	0.96	**0.85**	**0.90**	0.97	0.79	0.87	**0.98**	0.82	**0.9**
Hypercholesterolemia	**0.97**	0.43	0.60	0.96	0.53	**0.69**	**0.97**	0.50	0.66	0.96	**0.54**	**0.69**
Hypertension	**0.98**	**0.96**	**0.97**	0.97	**0.96**	**0.97**	**0.98**	0.94	0.96	0.97	**0.96**	**0.97**
Hypertriglyceridemia	**1**	0.18	0.31	**1**	0.21	0.35	**1**	0.18	0.31	**1**	**0.22**	**0.36**
OA	**1**	0.49	0.66	**1**	**0.57**	**0.73**	0.98	0.52	0.68	0.98	0.52	0.68
Obesity	**1**	0.9	0.94	**1**	**0.95**	**0.97**	**1**	0.75	0.86	**1**	0.85	0.92
OSA	0.94	0.69	0.79	0.94	0.70	0.80	0.94	**0.71**	0.81	**1**	**0.71**	**0.83**
PVD	**1**	**0.82**	**0.90**	0.99	0.79	0.88	0.99	0.79	0.88	0.96	**0.82**	0.88
Venous Insufficiency	**1**	**0.05**	**0.09**	**1**	0.02	0.04	**1**	**0.05**	**0.09**	**1**	0.02	0.04
Average	0.96	**0.61**	**0.70**	0.84	0.59	0.65	**0.98**	0.58	0.68	0.90	0.60	0.68

[b.] P is precision, R is recall, and F1 is F1-score. For each of the medical conditions, the highest values are in bold

To further evaluate the performance for different clinical entity extraction tools, we evaluate the performance by comparing the results from the four different clinical entity extraction tools with ground truth for each of the 15 clinical conditions which are linked to obesity (see Table I). As there is a difference between the two ground truth which were annotated by two different obesity experts, in this work we choose to use the Ground truth 1 as the ground truth. As shown in Table VII, for the averages, the CLAMP achieves the highest value on precision (0.98), and MetaMap achieves the highest value on recall (0.61) and F1-score (0.70).

From the results obtained in Table VI and Table VII, it is noted that on average MetaMap has a better performance compared with other tools. It is also noted that for a specific tool, it can achieve good performance in some clinical conditions but bad performance in other conditions. For example, cTAKES has the lowest average values in precision, recall and F1-score but it outperforms other tools in several clinical conditions such as Asthma, CAD, Diabetes, OA, Obesity, etc. Additionally, cTAKES is not able to identify any of the clinical conditions on Depression and Gallstones, which may be due to the configurations of the tool and the its abbreviation list.

A heatmap (Fig. 2) has been generated using the evaluation metrics from Table VII showing the performances for different clinical entity extraction tools on different medical conditions. To further explore the performance on different medical conditions for each of the clinical entity extraction tools used in this work, we have plotted the F1-score of each clinical entity extraction tools (Fig. 3). It can be seen that for each clinical entity extraction tools, the performance on extraction of different clinical conditions varied. For clinical conditions such as Hypertension, Obesity, OSA and PVD, all four clinical extraction tools have good

performances. However, for clinical conditions such as Venous Insufficiency, Hypertriglyceridemia, all four clinical extraction tools have performed poorly. It is also noted that the performance on extraction clinical conditions like Gallstones, CAD, depression, the performance of the four different clinical entity extraction tools varied significantly, which may due to the different built-in abbreviation lists of different clinical entity extraction tools.

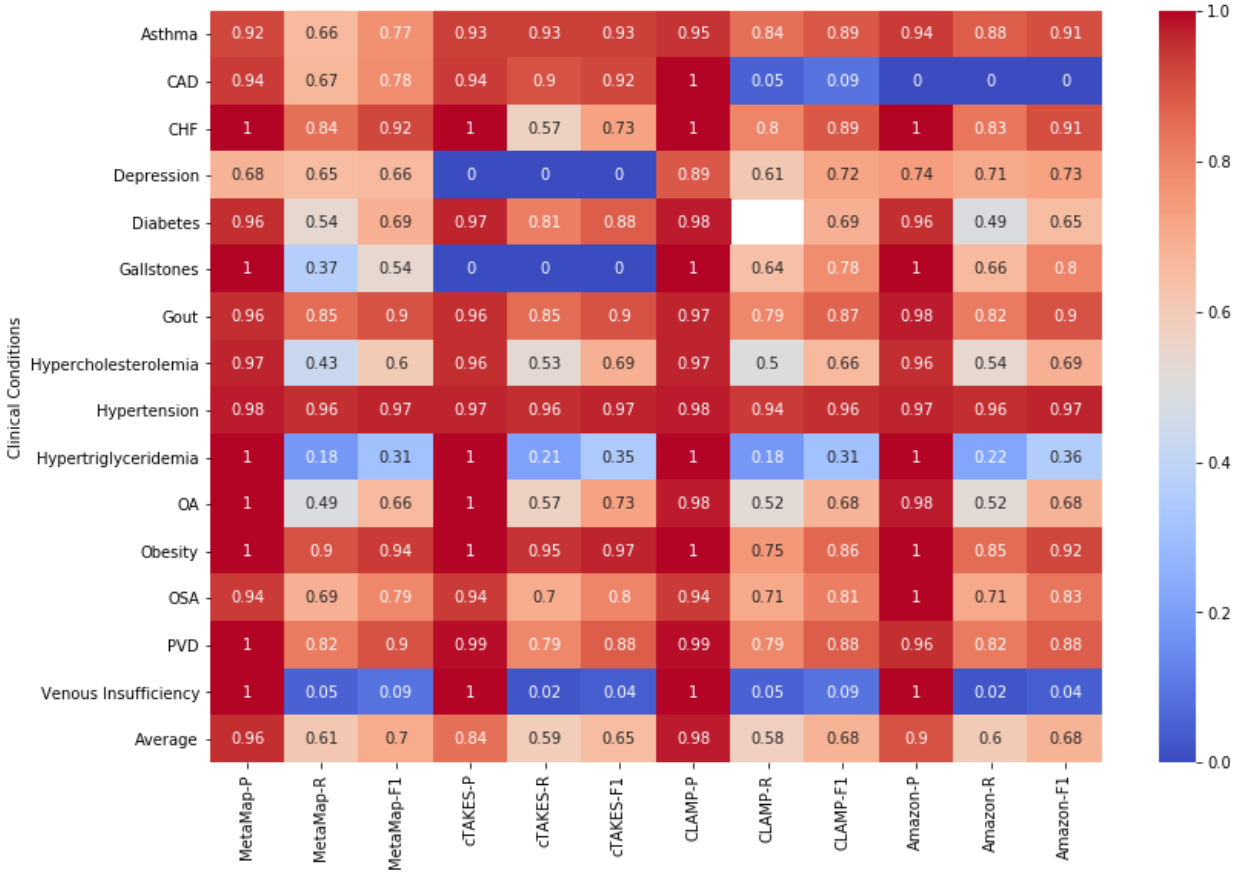

Fig. 2. Heatmap of the evaluation metrics for all four different clinical entity tools

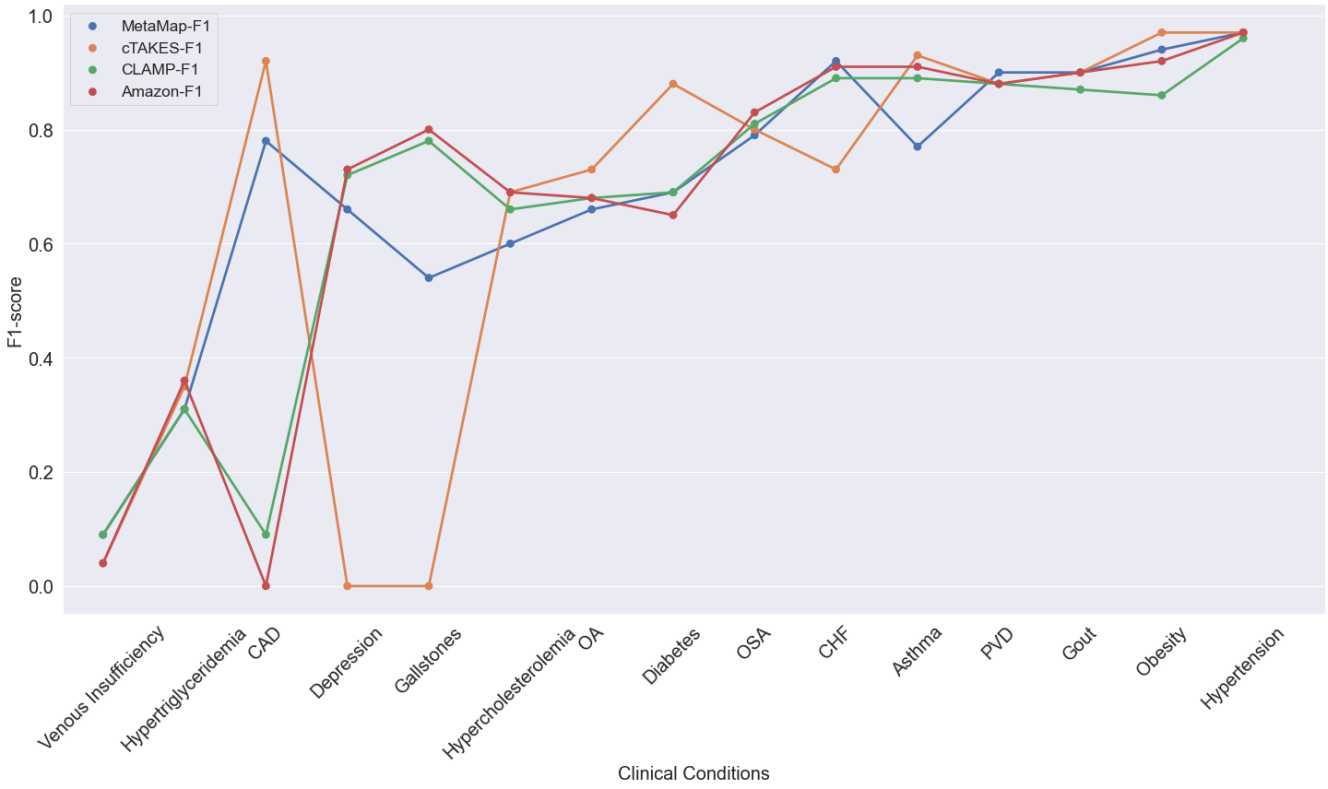

Fig. 3. Line plot of the F1-scores for each of the clinical entity extraction tools for different clinical conditions

V. Conclusion

In this paper, we have evaluated four automatic clinical entity extraction tools including cTAKES, MetaMap, CLAMP and Amazon Comprehend Medical API on the ib2b obesity dataset. For the results of the total number of the clinical conditions extracted, all the tools extract smaller numbers than the ground truth. cTAKES has slightly outperformed all other tools in the total number of clinical concepts extraction. Furthermore, we have evaluated the performance of all the extraction tools by comparing with the annotations from obesity experts. For the averages of all the selected clinical conditions, MetaMap has the highest recall (0.61) and F1-score (0.70), and CLAMP has the highest precision (0.98). The Amazon Comprehend Medical shows good performance for most of the clinical conditions except CAD. CLAMP also shows good performance for a number of clinical conditions. Although cTAKES outperforms other tools on several clinical conditions, it fails to identify Depression and Gallstones from all the clinical notes. In future, we will combine the results from different clinical entity tools to further improve the accuracy.

References

[1] R. S. Evans, "Electronic Health Records: Then, Now, and in the Future," Yearb. Med. Inform., 2016, doi: 10.15265/IYS-2016-s006.

[2] Y. Wang et al., "Clinical information extraction applications: A literature review," Journal of Biomedical Informatics. 2018, doi: 10.1016/j.jbi.2017.11.011.

[3] R. Reátegui and S. Ratté, "Comparison of MetaMap and cTAKES for entity extraction in clinical notes," BMC Med. Inform. Decis. Mak., 2018, doi: 10.1186/s12911-018-0654-2.

[4] M. Singh, A. Murthy, and S. Singh, "Prioritization of Free-Text Clinical Documents: A Novel Use of a Bayesian Classifier," JMIR Med. Informatics, 2015, doi: 10.2196/medinform.3793.

[5] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," Lingvisticae Investig. Investig. Int. J. Linguist. Lang. Resour. Investig. / Int. J. Linguist. Lang. Resour. Investig., 2007, doi: 10.1075/li.30.1.03nad.

[6] A. R. Aronson and F. M. Lang, "An overview of MetaMap: Historical perspective and recent advances," J. Am. Med. Informatics Assoc., 2010, doi: 10.1136/jamia.2009.002733.

[7] G. K. Savova et al., "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications," J. Am. Med. Informatics Assoc., 2010, doi: 10.1136/jamia.2009.001560.

[8] Y. Wu, J. C. Denny, S. T. Rosenbloom, R. A. Miller, D. A. Giuse, and H. Xu, "A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries.," AMIA Annu. Symp. Proc., 2012.

[9] T. Cai et al., "Natural language processing technologies in radiology research and clinical applications," Radiographics, 2016, doi: 10.1148/rg.2016150080.

[10] K. Denecke, "Extracting Medical Concepts from Medical Social Media with Clinical NLP Tools : A Qualitative Study," Int. Conf. Lang. Resour. Eval. (LREC 2014), 2014.

[11] A. Rodríguez-González, R. Costumero, M. Martinez-Romero, M. D. Wilkinson, and E. Menasalvas-Ruiz, "Extracting Diagnostic Knowledge from MedLine Plus: A Comparison between MetaMap and cTAKES Approaches," Curr. Bioinform., 2017, doi: 10.2174/1574893612666170727094502.

[12] E. Soysal et al., "CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines," J. Am. Med. Informatics Assoc., 2018, doi: 10.1093/jamia/ocx132.

[13] "Amazon Comprehend Medical." https://aws.amazon.com/comprehend/medical/.

[14] C. Liu et al., "Ensembles of natural language processing systems for portable phenotyping solutions," J. Biomed. Inform., 2019, doi: 10.1016/j.jbi.2019.103318.

[15] T. D. Imler, J. Morea, and T. F. Imperiale, "Clinical decision support with natural language processing facilitates determination of colonoscopy surveillance intervals," Clin. Gastroenterol. Hepatol., 2014, doi: 10.1016/j.cgh.2013.11.025.

[16] S. Gehrmann et al., “Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives,” PLoS One, 2018, doi: 10.1371/journal.pone.0192360.

[17] Ö. Uzuner, “Recognizing Obesity and Comorbidities in Sparse Data,” J. Am. Med. Informatics Assoc., 2009, doi: 10.1197/jamia.M3115.

[18] O. Bodenreider, “The Unified Medical Language System (UMLS): Integrating biomedical terminology,” Nucleic Acids Res., 2004, doi: 10.1093/nar/gkh061.

[19] L. T. T. Tran, G. Divita, M. E. Carter, J. Judd, M. H. Samore, and A. V. Gundlapalli, “Exploiting the UMLS Metathesaurus for extracting and categorizing concepts representing signs and symptoms to anatomically related organ systems,” J. Biomed. Inform., 2015, doi: 10.1016/j.jbi.2015.08.024.

[20] K. A. Spackman, K. E. Campbell, and R. A. Côté, “SNOMED RT: A Reference Terminology for Health Care,” J. Am. Med. Informatics Assoc., 1997.

[21] A. R. Aronson, “Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.,” Proc. AMIA Symp., 2001.

[22] D. Demner-Fushman, W. J. Rogers, and A. R. Aronson, “MetaMap Lite: An evaluation of a new Java implementation of MetaMap,” J. Am. Med. Informatics Assoc., 2017, doi: 10.1093/jamia/ocw177.

[23] “UMLS REST API Home Page.” https://documentation.uts.nlm.nih.gov/rest/home.html.