Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210160

Entity Extraction for Clinical Notes, a Comparison Between MetaMap and Amazon Comprehend Medical

Fatemeh SHAH-MOHAMMADI[a,1], Wanting CUI[a] and Joseph FINKELSTEIN[a]
[a]*Icahn School of Medicine at Mount Sinai, New York, NY, USA*

Abstract. Extracting meaningful information from clinical notes is challenging due to their semi- or unstructured format. Clinical notes such as discharge summaries contain information about diseases, their risk factors, and treatment approaches associated to them. As such, it is critical for healthcare quality as well as for clinical research to extract those information and make them accessible to other computerized applications that rely on coded data. In this context, the goal of this paper is to compare the automatic medical entity extraction capacity of two available entity extraction tools: MetaMap (MM) and Amazon Comprehend Medical (ACM). Recall, precision and F-score have been used to evaluate the performance of the tools. The results show that ACM achieves higher average recall, average precision, and average F-score in comparison with MM.

Keywords. MetaMap (MM), Amazon Comprehend Medical (ACM), Entity Extraction, Clinical documents.

1. 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 notes in the form of discharge summaries, radiology notes and progress notes). Although clinical notes with a narrative style and unstructured format provide a more complete image of patients' health information and condition, they complicate information extraction which is critical to computerized applications that rely on coded data in a daily healthcare setting, as well as to clinical research that leverages structured medication data [1]. Nowadays, various tools exist to extract information from clinical notes created in an unstructured format [2]. Two such tools are MetaMap (MM) and Amazon Comprehend Medical (ACM). The main difference between these tools is that MM is a rule-based while ACM is a neural network-based entity extraction tool. In addition, MM is amongst the most frequently cited medical entity extraction platforms.

Extracted medical information also form the basis for other tasks such as disease correlation, classification and diagnosis [3-5]. Due to the significance of medical entity extraction, this paper aims to compare the entity extraction performance of two tools (MM and ACM) with different computation approach. For this project, we worked with

[1] Corresponding Author, Fatemeh Shah-Mohammadi, Icahn School of Medicine at Mount Sinai, 1770 Madison Ave, 2nd Fl, New York, NY, USA, 10035; E-mail: fatemeh.shah-mohammadi@mountsinai.org.

the 2014 i2b2 NLP challenge data set for identifying Heart Disease and its Risk Factors in diabetic patients. The automated extraction resulted from MM and ACM was evaluated against the expert's manual annotations. We believe that this is the first work that compares the entity extraction performance of MM and ACM.

2. Methods

MM is well-known and widely used rule-based entity extraction tool in the biomedical field. This tool was developed by National Library of Medicine (NLP) in order to map biomedical text to concepts in the Unified Medical Language System (UMLS). To implement the mapping, MM uses a hybrid approach which combines a knowledge-intensive approach, natural language processing (NLP) and computational linguistic techniques [6,7]. Amazon Comprehend Medical has been 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 [8-10].

2.1. Dataset

The 2014 i2b2 heart disease and its associated risk factors identification dataset consists of 521 medical records with distribution of 8 disease risk factor categories and 38 associated indicators [11]. Due to the space limitation, we only considered 6 categories and some entities associated to each. The medical records in data set were in XML format where each record was composed of the actual narrative text note (corpus) and the annotations. We first separated the corpus from the annotations, and then imported the XML formatted annotations into a relational database to facilitate data analytics. It should be noted that annotations for every entity includes the original annotated text in the corpus and the position of the text in the corpus. We have considered 30 entities listed in Table 1 and Table 3. Table 1 shows the list of those entities identified by both tools. Column 2 and 3 in this table list the preferred name and UMLS Concept Unique Identifier (CUI) associated to each entity. The overlap between the output of MM and AMC enables the comparability of their performance. Both output not only the *text* for the extracted entity, but the boundaries (position) of the extracted entity in the corpus. We scored their performance based on the *text*, whether the predicted entity boundaries were correct.

Table 1. List of entities

Entities annotated by experts	Preferred name	CUI
Hypertension		
Hypertension	Hypertensive disease	C0020538
Hypertensive	Hypertensive (finding)	C0857121
htn	Hypertensive disease	C0020538
Hyperlipidemia		
Hyperlipidemia	Hyperlipidemia	C0020473
Dyslipidemia	Dyslipidemias	C0242339
Hypercholesterolemia	Hypercholesterolemia	C0020443
High Cholesterol	Hypercholesterolemia	C0020443

Diabetes		
Diabetes Mellitus	Diabetes Mellitus	C0011849
Diabetic	Diabetic	C0241863
DM	Myotonic Dystrophy 1	C3250443
Insulin Dependent Diabetes Mellitus	Diabetes Mellitus, Insulin-Dependent	C0011854
Non Insulin Dependent Diabetes Mellitus	Diabetes Mellitus, Non-Insulin-Dependent	C0011860
Obese		
Obesity	Obesity	C0028754
Morbid Obesity	Obesity, Morbid	C0028756
CAD		
Coronary Artery Disease	Coronary Artery Disease	C1956346
Coronary Artery Bypass Surgery	Coronary Artery Bypass Surgery	C0010055
Myocardial Infarction	Myocardial Infarction	C0027051
MI	Myocardial Infarction ECG Assessment	C3810814
Chest Pressure	Chest pressure	C0438716
Medication		
Zestril	Zestril	C0678140
Lipitor	Lipitor	C0593906
Verapamil	Verapamil	C0042523
Beta-Blocker	Adrenergic beta-Antagonists	C0001645

2.2. Evaluation Metrics

The experts' annotations have been considered as a gold standard to evaluate the automatic entity extraction of the two tools. We have used recall (or sensitivity), precision and F-score measured metrics to evaluate the results. Note that these metrics have been denoted as R, P and F in the result tables, respectively. For each entity, the scores for the three measured metrics have been calculated. Then, we averaged over the scores across all entities to calculate the average recall, precision and F-score achieved by the two tools. The selected programing language for all analysis was Python 3.8.

3. Results

The results have been shown in Table 2 and Table 3. Considering Table 2, the averages for the recall (R), precision (P) and F-score (F) with MM were 0.88, 0.83, and 0.82, respectively. With ACM, the averages for the same measures were 0.97, 0.86, and 0.90, respectively. In comparison with MM, ACM showed better performance by around 10% for the recall, 4% for the precision, and 10% for the F-score. MM showed a low recall value for hypertensive (0.29), beta-blocker (0.39), diabetic (0.51), and MI (0.55). Coronary artery bypass surgery presents a satisfactory recall value (0.72) although much lower than the overall results. Also, ACM had low recall values for coronary artery bypass surgery (0.57).

The Clinical notes contain many abbreviations, acronyms, and specialized terms that renders the extraction of patient information difficult. Table 3 shows the list of entities which exist in the data set but has not been identified by either MM or both tools. Based on Table 2, abbreviations such as "DM" and "htn" were identified by both tools, while as shown in Table 3 terms such as "severely obese", "insulin dependent diabetes", "insulindependent diabetes", and "insulin requiring diabetes" were not. According to Table 3, MM is sensitive to abbreviation used in clinical notes as it was not able to extract "high chol" which is abbreviation for high cholesterol. While the same word has been extracted by ACM with recall (R), precision (P) and F-score (F) equal to 1. In addition,

ACM has perfectly identified the terms “morbidly obese” and “increase cholesterol”, while MM has not.

Table 2. Summary of the evaluation

Entities annotated by experts and frequency of occurrences	Evaluation					
	MM			ACM		
	R	P	F	R	P	F
Hypertension (264)	1	0.74	0.85	1	0.93	0.96
Hypertensive (14)	0.29	1	0.44	1	0.68	0.76
htn (352)	1	0.78	0.88	1	0.8	0.89
Hyperlipidemia (166)	1	0.59	0.74	1	0.86	0.92
Dyslipidemia (24)	1	0.69	0.81	1	0.86	0.92
Hypercholesterolemia (3)	1	0.66	0.8	1	0.98	0.99
High Cholesterol (12)	1	0.67	0.8	1	0.92	0.96
Diabetes Mellitus (4)	0.75	1	0.86	1	1	1
Diabetic (17)	0.51	1	0.69	1	0.59	0.74
DM (268)	1	0.94	0.97	1	0.92	0.96
Insulin Dependent Diabetes Mellitus (1)	1	1	1	1	1	1
Non Insulin Dependent Diabetes Mellitus (1)	1	1	1	1	1	1
Obesity (70)	1	0.75	0.85	1	0.96	0.98
Morbid Obesity (13)	1	0.75	0.87	1	0.69	0.81
Coronary Artery Disease (104)	1	0.71	0.83	1	0.89	0.94
Coronary Artery Bypass Surgery (7)	0.72	1	0.83	0.57	1	0.73
Myocardial Infarction (41)	1	0.8	0.89	1	0.76	0.86
MI (68)	0.55	1	0.71	1	0.68	0.81
Chest Pressure (7)	1	1	1	1	0.47	0.63
Zestril (56)	1	0.53	0.76	1	0.81	0.9
Lipitor (201)	1	0.64	0.78	1	0.91	0.95
Verapamil (19)	1	0.79	0.88	1	1	1
Beta-Blocker (26)	0.39	1	0.56	0.77	1	0.87
AVERAGE	**0.88**	**0.83**	**0.82**	**0.97**	**0.86**	**0.90**

Table 3. List of unidentifiable entities

Tag name	Entities annotated by experts and frequency of occurrences	MM	ACM		
			R	P	F
Hyperlipidemia	High Chol (1)	nan	1	1	1
	Increased Cholesterol (1)	nan	1	1	1
Diabetes	Insulin Dependent Diabetes (1)	nan	nan	nan	nan
	Insulindependent Diabetes (5)	nan	nan	nan	nan
	Insulin Requiring Diabetes (1)	nan	nan	nan	nan
Obese	Morbidly Obese (7)	nan	1	1	1
	Severely Obese (2)	nan	nan	nan	nan

4. Discussion

Considering the results shown in Table 2, ACM resulted in better performance in comparison with MM with 10% higher average recall, 4% higher average precision, and 10% higher average F-score. In comparison to ACM, lower performance of MM in terms

of recall lies in the fact that MM is a rule and dictionary-based entity extraction tool. Rule-dictionary-based tools perform with high precision but low recall on the entity recognition tasks, showing a lack of generalization. Poor recall performance of these tools usually stems from their inability in identifying multi word phrases as concepts, unless exact matches can be found in the dictionary. Low frequent abbreviations such as "high Chol" and "MI" were also either not identified or identified with a low recall value by MM. It means that MM is sensitive to abbreviations in clinical notes. In addition, since ACM is a neural network-based tool, its training dataset included a wider range of vocabularies. As a result, out of 30 considered entities, ACM was able to identify 26, while MM has only identified 23. However, both tools have limitations in detecting misspelled words, missing words and spacing issues. In future studies, we will evaluate the entity extraction performance of more tools.

5. Conclusion

Majority of data in EHR are in the form of free text notes which feature gold mine of information. The information from these notes must be extracted and categorized to be utilized for clinical decision support, quality improvement and research. Therefore, an automated system will be necessary in order to parse medical information with high efficiency and accuracy. In this paper, we compared the automatic extraction of 30 entities associated with diabetes and heart disease using MM and ACM. Automatic extraction was compared to manual annotation by experts. The result of our conducted experiments on 23 entities listed in Table 2 proved that ACM outperforms MM by 10% for the average recall, 4% for the average precision, and 10% for the average F-score. In addition, ACM included a wider range of vocabularies and was able to identify higher number of entities (26 out of entire considered 30 entities in this paper). Based on our conducted analysis in this paper, we will proceed with ACM for real-world applications.

References

[1] Alnazzawi N, Thompson P, Batista-Navarro R, Ananiadou S, Using text mining techniques to extract phenotypic information from the PhenoCHF corpus. InBMC Med Inform Decis Mak. 2015;15(2):1-10.
[2] 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.
[3] Roque FS, Jensen PB, Schmock,et al.Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol. 2011 Aug;7(8):e1002141. Yildirim P, Çeken Ç, Hassanpour R, Tolun MR. Prediction of similarities among rheumatic diseases. J Med Syst. 2012; 36(3):1485–90.
[4] Bejan CA, Xia F, Vanderwende L, Wurfel MM, Yetisgen-Yildiz M. Pneumonia identification using statistical feature selection. J Am Med Inform Assoc. 2012 Sep-Oct;19(5):817-23
[5] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. InProceedings of the AMIA Symposium. 2001.p.17-21.
[6] Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. JAMIA. 2010;17(3):229-36.
[7] Bhatia P, Celikkaya B, Khalilia M, Senthivel S. Comprehend Medical: A Named Entity Recognition and Relationship Extraction Web Service, 18th IEEE ICMLA. 2019.p.1844-1851.
[8] Jin, M, Bahadori MT, Colak A, et al. Improving hospital mortality prediction with medical named entities and multimodal learning. arXiv preprint arXiv:1811.12276. 2018.
[9] Bhatia P, Arumae K, Celikkaya EB. Dynamic transfer learning for named entity recognition. InInternational Workshop on Health Intelligence. 2019.p.69-81.
[10] Stubbs A, Kotfila C, Xu H, Uzuner Ö. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. J Biomed Inform. 2015 Dec;58 Suppl(Suppl):S67-77.