Annotation of a Large Clinical Entity Corpus

Pinalkumar Patel
Team Lead, Research
ezDI Inc. Kentucky
pinal.p@ezdi.us

Disha Davey
Medical Coding Analyst
ezDI Inc. Kentucky
disha.d@ezdi.us

Vishal Panchal
Research Engineer
ezDI Inc. Kentucky
vishal.p@ezdi.us

Parth Pathak
Data Scientist
Microsoft
papatha@microsoft.com

Abstract

Having an entity annotated corpus of the clinical domain is one of the basic requirements for detection of clinical entities using machine learning (ML) approaches. Past researches have shown the superiority of statistical/ML approaches over the rule based approaches. But in order to take full advantage of the ML approaches, an accurately annotated corpus becomes an essential requirement. Though there are a few annotated corpora available either on a small data set, or covering a narrower domain (like cancer patients records, lab reports), annotation of a large data set representing the entire clinical domain has not been created yet. In this paper, we have described in detail the annotation guidelines, annotation process and our approaches in creating a CER (clinical entity recognition) corpus of 5,160 clinical documents from forty different clinical specialities. The clinical entities range across various types such as diseases, procedures, medications, medical devices and so on. We have classified them into eleven categories for annotation. Our annotation also reflects the relations among the group of entities that constitute larger concepts altogether.

1 Introduction

Corpus annotation is a process of adding interpretive linguistic information to a corpus (Leech, 2004). In the era of increasing trend of machine learning in NLP, annotated data drives the progress of NLP systems in many ways. Fields of NLP like Machine Translation, Information Extraction/Retrieval, Relationship Detection among the entities, depends heavily on an annotated corpus. In the past, many traditional NLP engines have used rule based dictionary lookup methods, with Unified Medical Language System (UMLS) (Lindberg et al., 1993) as a base dictionary, to detect clinical entities. However, these approaches fetch very low recall, due to the fact that the dictionary lookup can never capture all the lexical and linguistic variants of a medical term, and also due to the fact that the clinical documents depend upon a physician's writing style (Pathak et al., 2014).

Over the last few decades, Electronic Medical Records (EMR) have been an integral part of health care. Most of the data consists of a patient's symptoms, procedures being conducted and the medications prescribed to them. This data is mostly available in free text form or semi-structured form and may contain different level of difficulties in parsing this natural text and getting meaningful information. The extensive linguistic study is not available for a clinical domain to the extent it is for the general domain. Therefore, ML approaches are preferred over rule based approaches in the clinical domain. So a resource of the annotated corpus became a necessity to take full advantage of ML approaches. In the recent past, a few attempts have been made to annotate clinical texts. One of the first such attempt was made by (Wang, 2009) on 311 clinical notes from an Intensive Care Unit (ICU) department of the single hospital - Royal Prince Alfred Hospital (RPAH). However, no specific guidelines on how that data were annotated are available. Shared tasks like i2b2 in 2010 (Uzuner et al., 2011) (more than 800 documents), ShARe/CLEF (Suominen et al., 2013) (around 300 documents) in 2013 and SemEval 2014 (Pradhan et al., 2014), 2015 (Elhadad et al., 2015) (around 300 documents) have contributed a lot in increasing the availability of annotated clinical corpora. However, these corpora are focused only on certain type of entities. For example, i2b2 2010 data set has the annotation of three entity types, including test, treatment and disease.

There are other contributions to annotating clin-

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2033–2042
Brussels, Belgium, October 31 - November 4, 2018.

ical corpora, but depending on the purpose of the clinical research, most of the corpora generated in this domain are very specific to a disease or a disease category or some specialities of hospitals. For example, (Fiszman et al., 2000) annotated chest x-ray reports for automatic identification of acute bacterial pneumonia; (South et al., 2009) manually annotated clinical records to identify phenotypic information for inflammatory bowel disease; (Koeling et al., 2011) have annotated oncology reports on ovarian cancer for symptoms; (Xia and Yetisgen-Yildiz, 2012) have manually annotated the corpus for three different categories - Pneumonia (PNA), Critical Pulmonary Infection Score (CPIS) and critical recommendation on Radiology and ICU Reports; Clinical E-Science Framework (CLEF) corpus (Roberts et al., 2009) have annotated various types of clinical records from a single hospital, Royal Marsden Hospital, Oncology Center. This corpus is restricted to diagnosis, and only considers documents from the patients with neoplasms, that is only a primary diagnosis code in one of the top level sub-categories of ICD-10 Chapter II (neoplasms). As none of the above corpora cover the clinical domain in entirety, we have started building our own CER corpus.

In this paper, we demonstrate our approach of creating the corpus, deciding on annotation guidelines, annotation processes, annotation error analysis and improving the annotation quality with a corpus of 5,160 de-identified clinical documents for 11 different entity types varying from 40 different domains.

2 Corpus Creation

The success of many healthcare IT applications like computer assisted coding (CAC), clinical document improvement (CDI), core/quality measure monitoring are directly proportional to the accuracy of entity recognition and relations among the group of entities. Our aim of creating this corpus was to encapsulate as many entity types as possible, keeping in mind many of such future applications. We have annotated 5,160 de-identified clinical documents for 11 different entity types varying from 40 different domains. All the documents were de-identified using simple rule based approaches which follow safe harbour guidelines before any further use.

Clinical documents are very peculiar. Type of text in the document depends heavily on work types (like admit notes, discharge notes, operative notes, progress notes, etc.), associated medical domains (cardiology, oncology, endocrinology etc.) and varies considerably from physician to physician. So it was very important for us to make sure that we include as many different domains, work types and physicians as we could. So the first step towards corpus creation was to classify documents into different domains. We took around 700,000 documents, from 119 providers (hospitals and specialty clinics), unfortunately not all of these documents had the information regarding its domain and work type. We used a semi-automatic way to find domain related information and were able to classify 236,850 documents from all the documents into 40 different domains. We were also able to capture information regarding sub-specialities and physician expertise for these documents. Table 1 represents a sample domain classification.

Domain	**Docu--ment Count**	**Sub specialty**	**Physician**
Radiology	84635	2	16
Internal Medicine	15751	3	56
Emergency Medicine	12742	5	28
Cardiology	11555	6	40
Oncology	8325	4	11
Orthopedics	5480	2	19

Table 1: Domain wise classification of the documents

Work Type	**Document Count**
Consultation Reports	18
Operative Report	20
Progress Notes	40
History and Physical	15
Discharge Summary	30
Total	123

Table 2: Sample Document selection of gastroenterology domain

Based on this classification, our domain team prepared a list of important domain and worktype pairs from the whole corpus and on the basis of document distribution over these pairs, we filtered 5,160 documents to annotate which represent 40 different domains and more than 100 worktypes.

For example, Table 2 shows a sample selection of gastroenterology domain.

3 Annotation Guidelines

The first step towards preparing the annotation guideline was to decide on what has to be annotated. The best resource available to make this decision for the clinical domain was UMLS semantic group. The UMLS classifies the bio-medical entities into 133 semantic groups defined as Term Unique Identifiers (TUIs). These semantic groups are very fine-grained. We grouped multiple of these TUIs into different bucket to come up with 11 different types of medical entities. For the purpose of clinical information extraction, it is not necessary to use this much of detail as they would not be of much use in the clinical NLP applications, for example, clinical coding etc.

3.1 Clinical Entity Types

The clinical domain includes medical records like consultation reports, progress notes, history and physical, discharge summary, operative notes. These medical records comprise information about patient's diseases, affected anatomical area, procedures performed to treat the condition, devices used during the procedures and list of medications. It mainly covers the important entity types such as Problem, Anatomical structure, Procedure, Medical device and Medicine respectively. Apart from such information, the medical record also consists of the patient's normal functions, vital signs, lab examination and status of the patient. To cover this information, we have added other entity types Body function, Body measurement, Laboratory data and Finding respectively. Most of the times, entity type like Body measurement is mentioned with its values and to cover these values we have added an entity type named Measurement value which acts like a numerical modifier to add meaning to the entity type. Just a numerical entity has no meaning, so it is necessary that this entity type is always used in relation to an entity type named Body measurement. There are some words used in the clinical domain to add specificity to an entity. Such type of words are mostly adjectives and need a head word without which they have no contextual meaning and cannot be annotated alone. For example the word acute. The word acute alone has no specific meaning, but when it is relates to a head word entity like pain then it adds meaning to the word pain. To cover such information and to maintain the entities as simple and unique as possible, we have added an entity type named Modifier as a separate enhancing entity type. So in all, the gist of the medical record is captured by annotating the documents in our classified entity types.

Problem: The disease conditions which include major problem, disease, symptoms and disorders.

e.g.: Complication of **bleeding, infection, arterial puncture, DVT.**

Finding: Concepts apart from major problem, including abnormal conditions and the minor alteration in the regular condition.

e.g.: This is a 27-year-old female **gravida 4 para 1,** feeling **weak** and **lethargic.**

Procedure: Surgery or other procedures performed to cure or diagnose.

e.g.: This is an 82-year-old female with the history of **appendectomy,** status post **open reduction internal fixation.**

Anatomical Structure: Anatomical sites, cells and organs of the human body.

e.g.: The patient continued to have mild colitis throughout into the **cecum.**

Body Function: Activities carried out by the body to maintain the normal functioning.

e.g.: The patient's **breathing** was normal.

Lab Data: The type of analysis performed on blood, urine, other body substances or tissues to help diagnose or monitor the patient's condition.

e.g.: **AFB test** is negative, **TSH** is 1.1.

Body Measurement: The normal measurement of the body obtained without performing a complex procedure or test.

e.g.: The **weight** is up a couple pounds at 157 pounds, **pulse** is 74.

Measurement Value: Numerical value with its unit, associated with body measurement.

e.g.: The patient's heart rate was **90** and the BP was **140/90.**

Medical Device: Instruments used for the treatment, operation and various medical purposes.

e.g.: **Arterial line catheter** was placed over the **guidewire** without any resistance.

Medicine: A drug used for the treatment or prevention of a disease.

e.g.: Completed antibiotic course of **ceftriaxone.**

Modifier: Any word that adds some specific meaning to an Entity.

e.g.: **Chronic** skin excoriation due to known neurodermatitis.

We have classified our entity types in such a way that the meaning of the medical record is captured properly. Medical terminology is vast in nature, so other entity types except the 11 mentioned types are possible, like gene/variants, lipids, cells, cell lines, steroids, etc. and no doubt they are important entity types, but we do not need to annotate them separately because some of these entities can be classified in the existing entity types and others do not occur frequently in our medical records.

For example, an entity type like Lipid, documented only in laboratory data section is covered by our entity type Laboratory data.

e.g.: Cholesterol levels are high.

In this example, cholesterol is an organic lipid molecule present in the body and generally should be labelled under entity type Lipid. But in the medical records, it is mostly present in the form of a laboratory test. So there is no need to create a new entity type as it would be easily covered under our classified entity type named Laboratory data. We encounter such cases in our data and so entity type like Lipid is not used and if we use, it alters the meaning, which in this case, the system won't understand that it is a lab report and not an organic lipid molecule present in the body.

Another example is an entity type like Steroid, it is documented only in the medicine section which is covered by our entity type Medicine.

e.g.: The patient was prescribed corticosteroids.

"Corticosteroid" is a type of steroid hormone present in the body. But in this case, "corticosteroids" is a medicine which is prescribed and it is annotated under Medicine type. We encounter such cases in our data and entity type like Steroid is covered by our classified Medicine type.

So we have categorized these groups into broader categories of entity types like Problem, Procedure etc. These categories are as mentioned in Table 3 with relevant TUIs from UMLS.

Primary annotation guidelines were prepared by a linguist and an experienced medical coder. Both linguists and medical coder finalized entity types and created its initial descriptive definition with some complex examples. After that, both linguists and coder annotated the same set of documents separately and on the basis of conflicting annotations they sat together and solved the conflicts, then they updated the annotation guidelines to cover these conflict patterns. During this process, we have covered all possible domains and their existing document types in order to cover all the prominent entity patterns. After a few iterations, a well-defined annotation guideline was prepared and we started the actual annotation work. Note that the documents annotated during this process were not included in the corpus.

Entity Type	**Related TUI**
Problem	T046,T047,T048,T049, T050,T191,T037,T019,T184
Finding	T033,T034,T041,T084,T032, T201,T053,T054,T069,T068, T070,T067
Procedure	T059,T060,T061,T065,T058
Medicine	T200,T120,T110,T195,T131
Medical Device	T073,T074,T075,T203,T072, T071
Lab Data	T196,T119
Anatomical Structure	T017,T021,T023,T024, T025,T026,T029,T030
Body Function	T038,T039,T040,T042, T043, T044,T045

Table 3: Entity types with their mappings to TUIs

3.2 Inter-Conceptual Relationships

Relationships may exist between two or more concepts. In such instances, we considered inter-conceptual relationships between limited and frequently used terminologies. Relationships are marked only when relations occur in the same sentence and less frequent relationships like Lab data and problem, Medicine and Problem are rarely mentioned in a single sentence and hence we do not annotate those relations. We have not covered some relations like medicine with its attributes because it unnecessarily increases the complexity of the annotation task and these relations are easy to extract using finite state automata or regex based tools like cTAKES (Savova et al., 2010, 2011). Currently, the relationships are marked for the anatomical structure, measurement value and modifier with their related entities. Anatomical structures get related with three categories like problem, finding and procedure; measurement value gets related to only body measurement and modifier gets related to all other concepts except

measurement value.

3.2.1 Relationships with Anatomical Structures

Anatomical structure can be related to problem, procedure and finding. A relationship adds specificity to the concept.

Anatomical structure and problem relationship

This type of relationship helps to understand which part of the anatomical structure is affected by the particular problem. It simplifies the understanding of the problem and its area of effect.

For *e.g.*: she had severe pain in the left ankle.

Here pain is the problem and ankle is the anatomical structure. A relationship can be formed between these two concepts and it makes it easy to understand that pain is in the ankle.

Anatomical structure and finding relationship

This relation helps to understand to which anatomical structure is the finding related.

For *e.g.*: He noted he had felt some tingling and numbness in the left upper chest area.

Here tingling is the finding and chest is the anatomical structure. By linking these two concepts in a relationship we explain that tingling is occurring in the chest.

Anatomical structure and procedure relationship

This relationship explains that the procedure is being done at the given anatomical structure. It is usually used to relate any operation, test, and other such procedures to the organ or body site at which it is being performed.

e.g.: The patient had a CT of the brain.

Here CT is a procedure named Computed Tomography and brain is the anatomical structure. Relating these two concepts simplifies the understanding that CT of the brain was performed.

3.2.2 Relationships of Measurement Values

Measurement values cannot stand alone as only some numerical values do not provide any useful information. So these values need to be related to the concept where they belong to.

Body measurement and Measurement Value relationship

Body measurement is the measurement of basic body parameters like temperature, height, weight, pulse rate, etc. These parameters often have values that can be linked to the measurement in order to provide a complete meaning.

e.g.: Pulse: 80. BP: 110/70. Respirations: 16.

Here, pulse, BP, and respiration are the body measurements and 80, 110/70, and 16 are the measurement values of these respectively. So, forming the relationship between pulse and 80 signifies that the pulse rate is 80 per min, and similarly for all such concepts.

3.2.3 Relationships of Modifier

All the concepts except modifier itself and measurement values can be related to the modifier to add specific meaning to these concepts. Modifiers can never stand alone.

So after classifying the entity types to be annotated and the possible relationships between those entity types as mentioned above, a set of definite principles were documented. Annotators followed these principles while annotating the data. Some directive principles from the guideline are mentioned below.

3.2.4 Directive Principles

Consistency in the annotation is very important to get the good accuracy of systems using ML approaches. Therefore, to achieve good consistency, following simple rules were prescribed for the annotators:

- The entities should be annotated based on the above given entity types and also generate relationship as mentioned above.
- Modifiers and measurement values should not be annotated without any relationship with other entity types.
- Entities from section headers should not be annotated.
- Adjective should be annotated as a modifier only if it is not a part of an abbreviated term instead of the selected text. For example, in "Chronic Hypertension", 'Chronic' should be annotated as Modifier and 'Hypertension' should be annotated as Problem but in "Chronic Obstructive Pulmonary Disease" the whole phrase must be marked as Problem, as the phrase has a universally accepted abbreviation COPD. We refer to the UMLS dictionary for the abbreviated terms.
- Mark normal alternation in condition as a finding and abnormal alteration as a problem. i.e. "ST wave changes" is marked as a finding while "ST wave abnormality" as a problem.

- The entities should be annotated depending on the meaning and the context of the sentences because there might be some cases in which the type of an entity changes based on the sentence structure. For example: There was an evidence of "drainage" of pus and "drainage" was carried out. In this case, the word drainage appears twice, in the first half it means the pus is oozing out, so it is tagged as an entity type finding and in the second half it means a procedure named drainage was carried out and so tagged under entity type Procedure.

- There are instances where there are disjoint entities present in the sentence which have a relationship with other entities too. For example "The patient suffered from chest and abdominal pain." In this case, the word pain is a disjoint entity but it also has a relationship with the chest. So in such cases, we mark chest and abdominal as an anatomical structure respectively and pain as a problem and then generate a relationship between these entities.

4 Annotation Process

There are different ways to annotate a medical corpus, but we embarked on two pre-eminent approaches. The first approach is semi-automated annotation, in this approach, we can choose any pre-existing tools like MetaMap (Aronson, 2001), for CER task and thereafter annotator uses initial annotated results of tool to add, update or delete annotations as required and give the final annotated result. This approach definitely saves some human efforts and takes less time to annotate, however it is difficult to find an efficient and accurate tool. In addition, there is another issue with semi-automated annotation approach, that is, the annotator's decision may alter after viewing the output of the tool for ambiguous entities. The Second approach is to annotate the document based on just annotation guidelines and based on annotator's medical knowledge without looking at pre-annotated labels. As quality of annotation is more important than quantity of annotation specifically in medical domain, so we decided to go ahead with second approach.

Protégé (Noy et al., 2001) is an open source annotation and ontology editing tool. This tool was developed by Stanford Center for Biomedical Informatics Research. It allows us to add entity types with a description and assign color. Using this tool we can generate relationships between different entity types. The tool supports plain text, CSV and TSV format and so it is easy to load a medical document. The annotated data can be exported to a structured XML file format which makes the data widely accepted, machine readable and easy to use for any system.

Four annotators with a background in microbiology and biochemistry were involved in the annotation process. The medical document consists of major entities like the problem, procedure, anatomical structure and lab data which the annotators are well aware of because microbiology and biochemistry experts have a good knowledge about the anatomy and physiology as well as information about many ailments and procedures and laboratory tests. The queries were solved using UMLS dictionary and internal discussion with experienced medical experts and AAPC certified coders. The annotators are initially given a brief introduction about the importance of annotation and its applications and the impact of the annotated data on the clinical NLP. After that, the linguist and medical coders walk them through the set of guidelines which should be followed while annotating the corpus and present a demo of the annotating tool so that they get a brief knowledge of the tool. Initially, some documents were annotated in front of them so they become familiar with the process of annotation and tool. For the first few sets of documents (one set contains 30 documents), we divided the team of annotators into two pairs and made them annotate the same documents together. After completion of these sets, we changed the pairs of annotators. After 3 such iterations, each annotator had worked with all other annotators. This process was performed to make sure that all the annotators have the same level of understanding towards the task. The annotators are instructed to keep a note of problems faced while annotating data and discuss it with the medical coders regarding the medical annotation and with engineers regarding the technical difficulties and the functioning of the tool. While annotating ambiguous entity, UMLS metathesaurus was used as a reference. However, annotators found many inconsistencies in the UMLS. So in case of any disagreement among the annotators, the final decision about what to be annotated was taken by a

mutual consensus arrived after a discussion with the whole team. Note that the documents annotated during this process were not included in the corpus. After that, all the annotators were given a distinct set of documents in which 20% of the documents were the same between two annotators. We chose these same documents considering the distribution of worktypes over the corpus. These documents were used to calculate the kappa score (inter-annotator agreement). After annotation, the kappa score is calculated. The errors are discussed and solved by the annotators and then the data becomes useful. It took one year for the annotators to complete this task, in which the initial one month was spent on the training process.

4.1 Inter-annotator Agreement Calculation

Annotation is a mentally taxing task, and so annotators occasionally miss to annotate some of the medical entities, especially when a document contains a large number of them. The annotators are free to mark the entities according to their medical knowledge and as a result of that, some disagreement arises. An inter-annotator agreement is an important quality measure. The Cohen's kappa coefficient is used to estimate the agreement between the annotators. The Cohen's kappa equation is calculated with the following formula:

$$K = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e} \quad (1)$$

where, p_o is the relative observed agreement among annotators, and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the annotators are in complete agreement then = 1. If there is no agreement among the raters other than what would be expected by chance (as given by pe), $K \leq 0$ (Koeling et al., 2011). The initial kappa score was 68.01% calculated on the set of 30 documents. The aim was to achieve good kappa score up to 95% to minimize the conception gap between the annotators. The kappa score is obtained after each set of annotation and the annotators are made to sit together and review the whole document with their own annotation results. They internally discuss why some entities are annotated and missed or ambiguously annotated. After discussion, they come to a conclusion and make a note of the changes which helps to decrease the conception gap. Using these notes, they are clear about what to annotate and what not to. Using structured guidelines and proper classification of entity types and relations, we achieved a 96.89% kappa score for the entity and relationship annotation.

5 Corpus Statistics

Using the mentioned process, we have created a large annotated corpus of the medical domain for clinical entity and relationships which covers 5,160 clinical documents with 398,568 sentences. Out of these sentences, 190,188 sentences contained one or more medical entities which were annotated as concepts in the corpus, hence concept density over sentences is 47.72%. These 398,568 sentences vary in length from 1 token to 150+ tokens as shown in Figure 3. The corpus has average 9.59 tokens per sentence with total 3,825,465 tokens across the corpus. Out of these 3,825,465 tokens, 600,550 tokens annotated as concepts in the corpus, hence concept density over token is 15.70%. An EMR record consists of a gist of the patient's conditions, procedure and tests carried out, medications prescribed etc. But, along with that, there is a lot of insignificant information present in the record like hospital name and address, patient's demographics which are not healthcare entities. Apart from all these, an EMR contains information documented using definite templates and so a lot of sentence are generally not important and common in every medical record and they reflect in annotated token density. The frequency of annotation for each relation and entity type are detailed in Figure 1 and 2 respectively. There are in total 443,328 annotated concepts in the corpus. These annotated concepts result in an average length of 1.35 tokens per concept. The highest frequency concepts are Problem, Procedure, Medication, Anatomical Structure and Modifier which account for 78.81% of the data. The remaining 21.19% concepts are distributed into 6 rare categories. For the relation annotation, the corpus contains total 119,968 relations, out of which modifier - anatomical structure cover 30,018 relations, modifier with other entity types except anatomical structure cover 49,677 relations, anatomical structure with other entity types except modifier cover 35,401 relations and body measurement - measurement value cover 4,872 relations.

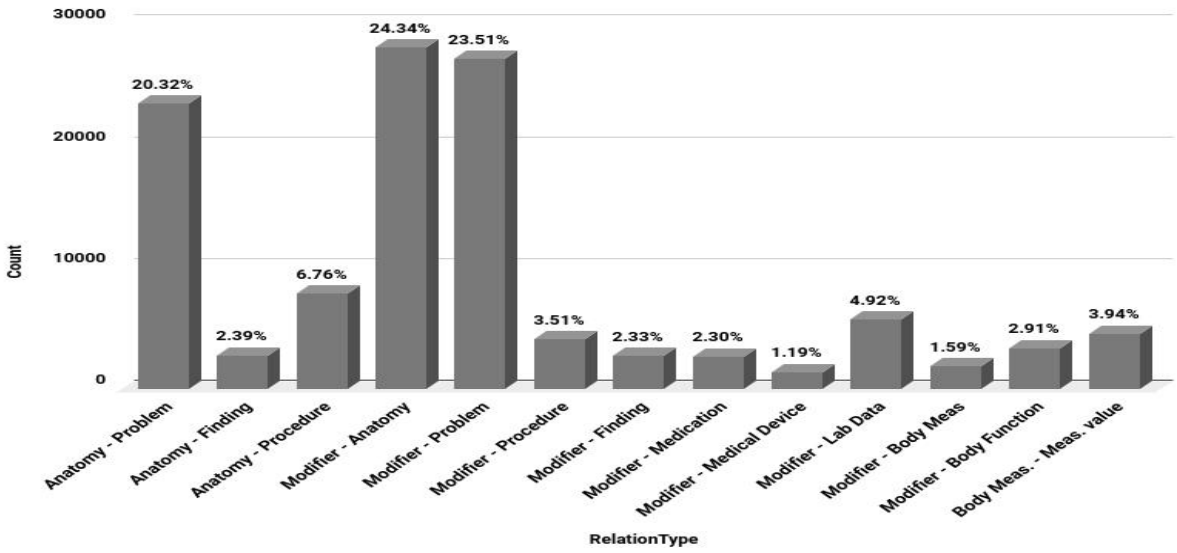

Figure 1: Frequencies of relation types

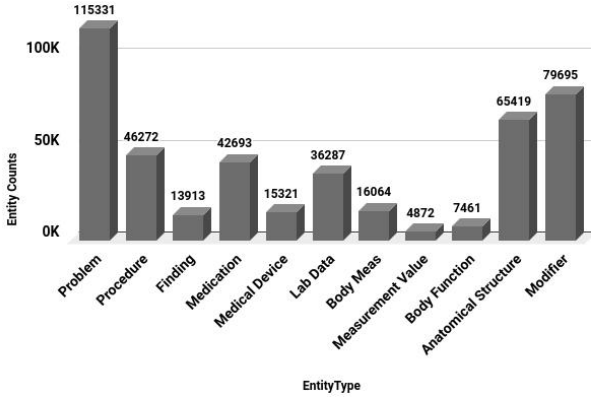

Figure 2: Frequencies of entity types

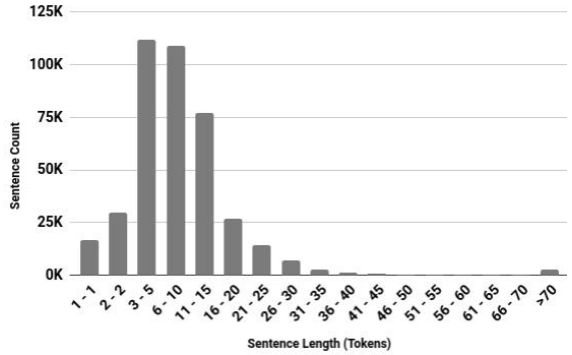

Figure 3: Sentence distribution over token counts

5.1 Initial Experiment with CRF

Initially, we used Conditional Random Fields (CRF) which is a well-known and a proven approach to detect continuous entities using CRF++ (Kudo, 2010) toolkit. Feature selection is the key task for accurate CRF model. We used feature sets as mentioned below (Pathak et al., 2015)

- Bag of words features, prefix, suffix
- Orthographic features, binary (true/false) like whole word capital, first char capital, numeric values, dates, words contains hyphen or slash, medical units (mg, gram, ltr, etc)
- Grammatical features like
  - Parts of Speech (PoS) (Choudhary et al., 2014), chunk, consistency parser (Oinam et al., 2018); all developed and designed in-house for clinical NLP
  - Head of the phrases
  - Stemming of the words
- Dictionary features, binary (true/false) was used to check whether the word is present in the medical dictionary or not.
- Stop words and word embedding id from word2vec trained on 2 lakh clinical documents
- Section header and document type information and sentence cluster id

Table 4 shows the initial accuracy, calculated using the Perl script provided by the CoNLL 2000 task (Tjong Kim Sang and Buchholz, 2000), on CRF using a BIO format with the combination of different features on 5-fold validation. We are able

Feature Set	Precision	Recall	F-Measure
Token	92.32	86.00	89.05
Orthographic	92.22	86.26	89.30
POS and Chunk	91.92	87.94	89.89
Dictionary	91.91	87.43	89.61
Consistency Parse	92.04	87.32	89.62
POS-Chunk and Orthographic	91.91	88.04	89.93
POS-Chunk and Dictionary	91.82	88.77	90.27
POS-Chunk, Consistency Parse and Dictionary	91.65	88.90	90.25
Above + Stemmer, Word Embedding Id	92.00	90.26	91.12
All	92.12	91.05	91.58

Table 4: CER task accuracy using CRF with different feature set

to get 91.58% f-measure with 92.12% precision and 91.05% recall using all features. We obtained a good performance using CRF which shows the quality of the annotation over the definite corpus.

6 Conclusion

In the domain of clinical NLP, there is a paucity of good and large annotated corpus. Research shows that some amount of data has been annotated according to different purposes and applications. Available annotated corpora are not found covering the clinical domain in entirety. In this paper, we have described the creation of a corpus of the clinical domain annotated with clinical entities and their inter-conceptual relationships. We have manually annotated the clinical corpus comprising of 5,160 documents, 398,568 sentences according to the guidelines created in-house.

Acknowledgements

We acknowledge the contribution of former annotators namely Meera Panchal, Namrata Srivastava, Yvonne Christian and Pranita Joshi, who provided insight and expertise that greatly helped in building the corpus. We would also like to thank Sagar Soni for his assistance with the functioning of the protégé tool.

References

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Narayan Choudhary, Parth Pathak, Pinal Patel, and Vishal Panchal. 2014. Annotating a large representative corpus of clinical notes for parts of speech. In *Proceedings of LAW VIII-The 8th Linguistic Annotation Workshop*, pages 87–92.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. Semeval-2015 task 14: Analysis of clinical text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310.

Marcelo Fiszman, Wendy W Chapman, Dominik Aronsky, R Scott Evans, and Peter J Haug. 2000. Automatic detection of acute bacterial pneumonia from chest x-ray reports. *Journal of the American Medical Informatics Association*, 7(6):593–604.

Rob Koeling, John Carroll, Rosemary Tate, and Amanda Nicholson. 2011. Annotating a corpus of clinical text records for learning to recognize symptoms automatically.

Taku Kudo. 2010. Crf++: Yet another crf toolkit (2005). *Available under LGPL from the following URL: http://crfpp. sourceforge. net*.

Geoffrey Leech. 2004. Developing linguistic corpora: a guide to good practice adding linguistic annotation. *Lancaster University*.

Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Methods of information in medicine*, 32(04):281–291.

Natalya F Noy, Michael Sintek, Stefan Decker, Monica Crubézy, Ray W Fergerson, and Mark A Musen. 2001. Creating semantic web contents with protege-2000. *IEEE intelligent systems*, 16(2):60–71.

Nganthoibi Oinam, Diwakar Mishra, Pinal Patel, Narayan Choudhary, and Hitesh Desai. 2018. A treebank for the healthcare domain. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 144–155.

Parth Pathak, Pinal Patel, Vishal Panchal, Narayan Choudhary, Amrish Patel, and Gautam Joshi. 2014. ezdi: A hybrid crf and svm based model for detecting and encoding disorder mentions in clinical notes. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 278–283.

Parth Pathak, Pinal Patel, Vishal Panchal, Sagar Soni, Kinjal Dani, Amrish Patel, and Narayan Choudhary. 2015. ezdi: a supervised nlp system for clinical narrative analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 412–416.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics*, 42(5):950–966.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Guergana K Savova, Janet E Olson, Sean P Murphy, Victoria L Cafourek, Fergus J Couch, Matthew P Goetz, James N Ingle, Vera J Suman, Christopher G Chute, and Richard M Weinshilboum. 2011. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. *Journal of the American Medical Informatics Association*, 19(e1):e83–e89.

Brett R South, Shuying Shen, Makoto Jones, Jennifer Garvin, Matthew H Samore, Wendy W Chapman, and Adi V Gundlapalli. 2009. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. In *BMC bioinformatics*, volume 10, page S12. BioMed Central.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 212–231. Springer.

Erik F Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task: Chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7*, pages 127–132. Association for Computational Linguistics.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Yefeng Wang. 2009. Annotating and recognising named entities in clinical notes. In *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, pages 18–26. Association for Computational Linguistics.

Fei Xia and Meliha Yetisgen-Yildiz. 2012. Clinical corpus annotation: challenges and strategies. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*.