DEGREE PROJECT IN COMPUTER SCIENCE AND ENGINEERING, SECOND CYCLE, 30 CREDITS

STOCKHOLM, SWEDEN 2017

Annotating Mentions of Coronary Artery Disease in Medical Reports

Annotation vid omnämnanden av kranskärlssjukdom i medicinska rapporter

LUKE TONIN

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF TECHNOLOGY AND HEALTH

Abstract

This work demonstrates the use of a machine learning text annotator to extract mentions or indications of coronary artery disease in unstructured clinical reports. The overall results prove the effectiveness of the technologies used and the possibility to use machine learning annotators for clinical information extraction purposes.

Keywords: Natural language processing, information extraction, coronary artery disease, watson knowledge studio

Acknowledgements

During my time with IBM, I was given the opportunity to develop the skills required for this thesis. Among other things, I improved my knowledge of machine learning, natural language processing, information extraction and coronary artery disease. I would like to thank Emmanuel Vignon, my IBM tutor who gave me to opportunity and freedom to search and explore all IBM products and knowledge sources. I also thank Pawel Herman who was my supervisor at KTH, Dmitry Grishenkov who led the master thesis course, and my classmates at KTH who read and gave feedback on my work.

Nomenclature

NLP	Natural Language Processing
IBM	International Business Machines
WKS	Watson Knowledge Studio
WCA	Watson Content Analytics
CAD	Coronary Artery Disease
I2B2	Informatics for Integrating Biology and the Bedside
ShARe	Shared Annotated Resources
LAD	Left Anterior Descending Artery
UIMA	Unstructured Information Management Applications
SaaS	Software as a Service
RCA	Right Coronary Artery
D1	First Diagonal Artery
SME	Subject Matter Expert
UMLS	Unified Medical Language System
NSTEMI	Non-ST Segment Elevation Myocardial Infarction
STEMI	ST Segment Elevation Myocardial Infarction
CABG	Coronary Artery Bypass Grafting

TABLE OF CONTENTS

1 INTRODUCTION

1.1 BACKGROUND

Unstructured data refers to information that is not represented in a predefined model. Unstructured data is usually useless to computer systems and requires humans to analyze. In healthcare, this refers to clinical narratives, scans, images, reports etc. The amount of unstructured clinical data is increasing. However, the capacity of medical professionals to read and analyze this data is constant and unstructured clinical data increasingly remains unused. In this thesis, unstructured data will refer to text data.

1.2 PURPOSE

If the knowledge contained in unstructured form was systematically analyzed, the extracted information could help detect weak signals or trends that would otherwise be invisible. Insights could be derived from the analysis of big data in the form of unstructured text. For instance, large amounts of clinical notes could be analyzed to detect correlations between a certain type of medication (annotated in the clinical notes) and the presence of coronary artery disease (CAD) that were previously unknown. Although the correlation would certainly not provide sufficient proof, it could give indications to researchers about where to search. A similar use case would be using the annotator to find negative correlations between a type of medication and a pathology. A negative correlation could indicate unknown positive effects of a certain drug. All applications have in common that they use the data extracted from sources that were previously unused by computer systems (unstructured text) to provide statistical insights or knowledge.

This paper will present the use of an IBM software named Watson Knowledge Studio to produce a machine learning text annotator to detect mentions of coronary artery disease in unstructured clinical reports.

2 METHOD

This section will present the source of the clinical documents used for training and testing. It will also describe two broad types of annotators, and the type of annotator chosen for this task. Finally, a specific description of the information extracted from the documents will be presented.

2.1 DATA SOURCE AND OBJECTIVE OF THE ANNOTATOR

To create an annotator, a large quantity of data is required. This was a major issue and still today remains one of the biggest hurdles for creating machine learning annotators in the medical field. Indeed, most clinical data is private and protected [1]. In the United States of America for instance, all medical data falls under the HIPAA policies that protect the privacy and security of individually identifiable health information [2]. These restrictions mean that most clinical data is inaccessible for research purposes and must at least be deidentified. In the past years, with the advent of automatic deidentification technologies large clinical datasets have become available for research purposes (IB2B, ShARe) [3] [4].

This thesis uses a set of clinical documents provided by I2B2 at a 2014 competition for annotating clinical narratives to detect risk factors for heart disease in diabetic patients [5]. In total, 14 papers were submitted for the annotation of heart disease risk factors. The aim of the competition was to detect heart disease risk factors as well as mentions of coronary artery disease. The annotation guidelines referred to 8 broad categories of annotations: diabetes, hyperlipidemia-hypercholesterolemia, hypertension, obesity, family history of CAD, smoking, medication and mentions of CAD. An analysis of the papers revealed good results for almost all categories. Interestingly, the hardest elements to detect were the mentions of CAD. For that reason, this paper will investigate the detection of mentions of CAD.

The objective of the annotator is to obtain the highest detection rate of the mentions of CAD on a test set of clinical documents.

2.2 TYPES OF ANNOTATORS

There are two broad categories of annotators: rule-based and machine learning. Rule based information extraction systems use explicit rules to extract information from text. Rule based methods do not require any training documents but require someone to spend time thinking of language rules that would help identify the concepts that are to be extracted.

Example: The patient had a blood pressure of 120/80 mmHg.

A rule could be formulated as follows: if the form "number/number" followed by "mmHg" is detected then annotate "number/number mmHg" with the tag "blood pressure".

Rule based methods are useful and effective when the entities to be extracted are in a structured form within the text or when the language rules can be easily formulated explicitly.

Watson Content Analytics is a IBM product that uses rule based methods to annotate text. It allows users to define their own ontology and therefore adapt to specific use cases. It is based on the UIMA (Unstructured Information Management Applications) pipeline [6].

The machine learning annotator used in this paper is based on supervised learning. It requires training data composed of text and the associated annotations. In the case of machine learning annotators, the user does not need to explicit the annotation rules but must, on the other hand, provide an annotated learning corpus that is injected into the algorithm to train the model.

Watson Knowledge Studio is a product developed by IBM. It is a development environment in which one can upload documents, manually annotate them, and use the data to train a machine learning model that will learn from the manually annotated documents and that can be used to provide automatic information extraction. Watson Knowledge Studio is a cloud based SaaS (Software as a Service) meaning that that the machine learning is done in the cloud and not on a local machine.

Rule base methods were the first to be developed and remain the basis for most industrial information extraction systems [7]. Machine learning methods are more recent and attract most of the attention in academia. The following table summarizes that pros and cons of both methods.

Table 1 Pros and Cons of Rule Based and Machine Learning Annotation Methods

	PROS	**CONS**
RULE BASED	- Rules are explicit - Easy to understand - Easy for non-experts to maintain and improve - Easy to adapt and add new rules - Easy to debug - Deterministic	- Limited capabilities, cannot integrate the subtleties of natural language - Requires time to think of all the rules
MACHINE LEARNING	- Can be trained with examples - Can detect elements through subtle implicit rules - Probabilistic as opposed to deterministic	- Often (although not always) requires ML expertise - Rules are implicit so the models are hard to debug - Requires a lot of planning

The limitations of rule based models appear when there are many subtly different ways of expressing the same thing. Due to the deterministic nature of the rules, even a subtle difference in the formulation or a typo can prevent the annotation.

The variety of ways in which CAD can be expressed in a clinical note suggests that the best way of improving on the state of the art for this task was to use a machine learning model.

2.3 EXTRACTED INFORMATION

The aim of the annotator is to automatically annotate mentions of CAD in clinical notes. Deciding what qualifies as a marker of CAD is not trivial and requires medical knowledge.

The indicators for CAD were split into 4 broad categories.

2.3.1 Direct Mentions of CAD

This category groups all direct mentions of CAD in the patients report a well as history of CAD. Mentions of conditions that indicate coronary artery disease were also added to this category. For instance, coronary arteriosclerosis or coronary ischemia are direct mentions of CAD in the patient. Although the surface form (the form in which these are written in the text) can vary (e.g. h/o instead of history of) direct mentions of CAD are usually easy to detect through machine learning.

2.3.2 Events

This category includes events in the life of the patient that would indicate the presence of CAD, these events include myocardial infarctions, medical procedures, or interventions. An example of an event could be "non-Q wave MI" which is a type of myocardial infarction that indicates CAD.

2.3.3 Test Results

This category includes all test results indicating the presence of CAD, these tests could be a catheterization procedure showing coronary stenosis (e.g. "LAD 50% lesion"), positive stress test results (e.g. "Stress (3/88): rev. anterolateral ischemia").

2.3.4 Symptoms

The final category includes all symptoms of CAD. The most common being chest pain (e.g. "mid-sternal chest discomfort", "substernal pain like a 'ball' pressing in on her chest").

These four categories of mentions of CAD are summarized in the following table.

Table 2 Indicators of CAD

INDICATOR	DESCRIPTION
Mention	A diagnosis of CAD, or a mention of a history of CAD
Event	- MI, STEMI, NSTEMI - Revascularization procedures (CABG, percutaneous) - Cardiac arrest - Ischemic cardiomyopathy
Test result	- Exercise or pharmacologic stress test showing ischemia - Abnormal cardiac catheterization showing coronary stenosis (narrowing)
Symptom	- Chest pain consistent with angina - Abnormal heart rhythm (arrhythmia)

2.4 Creation of the Model

The following section will present the steps necessary to create the annotator using Watson Knowledge Studio.

2.4.1 Creation of the Type System

The type system is composed of two elements: the groups of entities that are to be extracted from the text, and the relations between the entities. The model developed for this thesis contains the four groups of entities above: Mention, Event, Test Result and Symptom as well as a fifth entity that provides additional information about the Rate of an event or a test mentioned in the medical record. The relation contained in this model is the relation "dateOf" between the entities Test Result and Date, and Event and Date.

Following is a graph representation of the type system of the information extraction model.

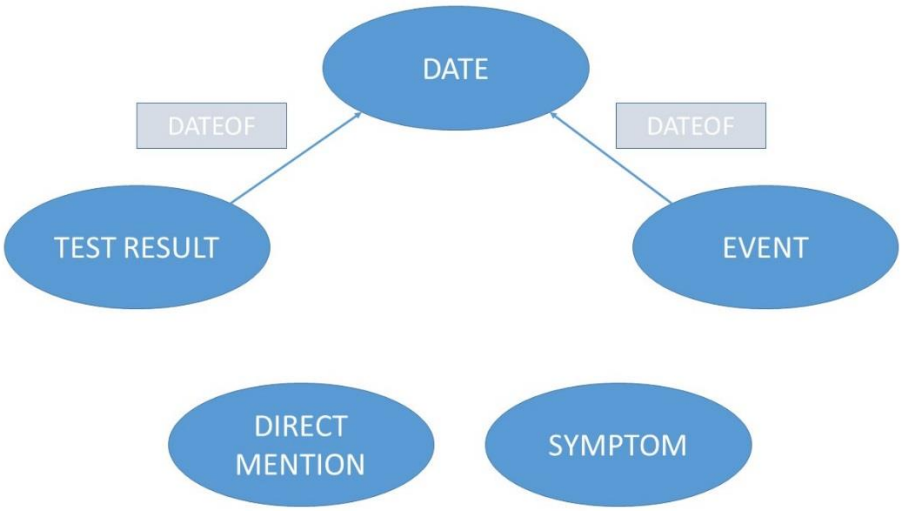

Figure 1 Graphical representation of the Type System

The type system is the structure of the information that is extracted from the clinical data. Following is an example of how the clinical data is annotated. There are three types of entities in this example: Mention (CAD), Date (1/73) and Test Result (80% prox RCA). There are also three mentions of the relation type dateOf which indicate that the 80% blockage of the RCA due to stenosis with thrombus and the D1 stenosis are from a test that took place in "January 2073".

Note: The documents have been deidentified, one task of the deidentification is to modify the dates since dates could provide information on the identity of the patient. In this set of documents, the dates in the clinical notes were modified to be in the future (January 2073 for instance).

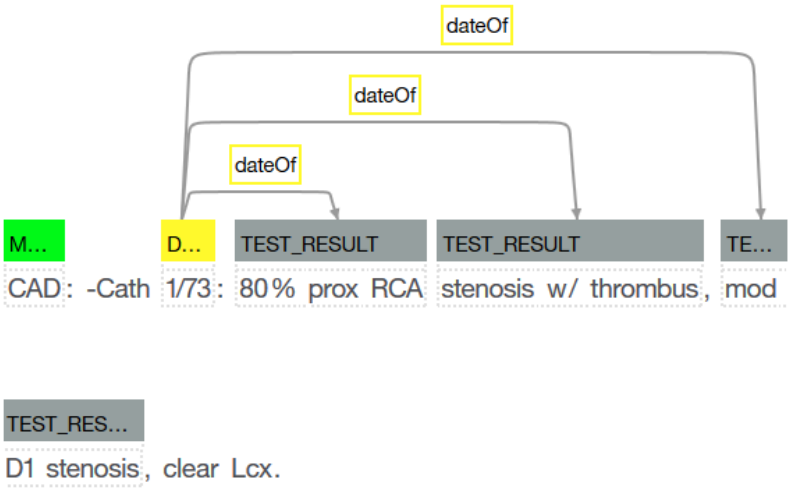

Figure 2 Example of the annotation of a clinical document

2.4.2 Manual Annotation of the Training data

Supervised machine learning requires labeled training data that it used to teach the machine. The amount of annotated data required to obtain satisfactory results varies greatly depending on the complexity of the type system and the language used in the documents. Manual annotation is not a task to be underestimated. For a given type system, there are many ways to annotate a single document. Human produced annotations will depend on the interpretation of both the type system and clinical notes. For instance, "chest pain and pressure with some shortness of breath after she walks about one block" could be annotated as the result of a stress test (therefore annotated with the entity Test Result) or a symptom of CAD (chest pain being a common symptom of CAD) (therefore annotated with the entity Symptom). Manual annotation requires expertise. It is not possible to identify in advance all the mentions of CAD that could appear in a text, therefore an annotator necessarily must judge whether a mention should be annotated or not. This is especially true in the medical field; manual annotators should be subject matter experts (SME).

Another difficulty appears when more than one person is creating the training data. Because of the subjective nature of the annotation task, two people are likely to annotate the documents slightly differently. The machine will have a harder time learning the annotation patterns if the training data is heterogeneous. Making sure that all annotators follow the same annotation guidelines is essential.

These difficulties have many consequences for the use of machine learning annotators in real world settings and are some of the reasons why rule based information extraction is still the most widely used.

In the case of this paper, the type system was relatively simple and a single person did all the annotations so many of the difficulties of manual annotation were avoided.

2.4.3 Evaluation and Reannotation

After producing enough annotated documents, they are used to train the machine learning model. The trained machine learning model is then evaluated. On the first evaluation, the model was not

accurate. However, it was sufficiently accurate to be used to annotate other documents; the annotations were then reviewed and used to retrain and improve the model. This iteration (using the annotator to help create training data) was done several times and allows to bootstrap the training of the annotator.

3 RESULTS

The following sections will present the method and metrics used to evaluate the annotator. It also describes the results for each metric and for each category of annotation (mention of CAD, CAD related events, symptoms, test results and dates).

3.1 EVALUATING THE ANNOTATOR

Automated annotation is an information extraction task and a common way of evaluating information extraction systems is to use precision and recall [8].

Calculating precision and recall requires some of the manually annotated documents to be used not as training documents but as test documents. To test the model, a certain fraction of the clinical documents is removed from the training data (30% for instance). The model is then trained with the remaining 70%. The removed 30% are then automatically annotated by the trained model and the annotations are compared with the manual annotations.

Two metrics were derived from this comparison: precision and recall.

Precision indicates whether the annotations detected by the annotator tend to be correct. This number is calculated by dividing the number of correct annotations retrieved by the number of retrieved annotations. In the best case, all retrieved annotations are relevant and the precision is 1. In the worst case, none of the retrieved annotations are relevant and the precision is 0.

$$PRECISION = \frac{CORRECTLY\ DETECTED\ ANNOTATIONS}{ALL\ DETECTED\ ANNOTATIONS}$$

Recall indicates whether the annotator annotates sufficiently. It is calculated by dividing the number of correct annotations by the total number of correct annotations. In the best case, all annotations that should be annotated are detected, and the recall is 1. In the worst case, none of the annotations that should be annotated are detected and the recall is 0.

$$RECALL = \frac{CORRECTLY\ DETECTED\ ANNOTATIONS}{TOTAL\ NUMBER\ OF\ CORRECT\ ANNOTATIONS}$$

Precision and recall can be combined to provide a single measure of performance called F1 score.

$$F1 = \frac{2 * PRECISION * RECALL}{PRECISION + RECALL}$$

3.2 TEST RESULTS

The annotator was tested using the methods described in the previous section.

The results are summarized in the following table:

Table 3 Results from annotator testing

ENTITY TYPE	F1 SCORE	PRECISION	RECALL
MENTION_CAD	0.83	0.81	0.85
EVENT	0.64	0.89	0.5
SYMPTOM	0.8	1	0.67
TEST RESULT	0.65	1	0.48
DATE	0.59	1	0.42
OVERALL	0.73	0.91	0.61

4 DISCUSSION AND CONCLUSIONS

4.1 DISCUSSION

At the time the results were extracted, the annotator had been trained on 60 clinical reports. This is sufficient to produce reasonable results but the quality of a machine learning annotator, like any supervised learning system, is highly dependent on the quality and quantity of the training data. For the results to improve, more training documents would have to be produced. For the model to reach its full potential, it would have to be trained with a couple of hundred documents. An analysis of the precision and recall reinforces the assumption that the quality of the annotator would improve with more training data. Indeed, the overall precision is of 91% and the overall recall is of 61%. Meaning that the algorithm is accurate (that is doesn't often detect things that it should not have), but it is not exhaustive in that it does not detect all the mentions that it should be detecting. By increasing the training data, the model will have more examples of the annotations and will not miss them as much.

This test shows that all annotations do not perform the same. Surprisingly, the date tag possesses the lowest F1 score due to a very low recall (42%). This is possibly because there were less mentions of the date in the training text and that they appeared in very different surface forms (12/2084, 21/12/2084, December 2084 and 21st of December 2084). The model could have encountered difficulties because the compact date and the expanded date look so different. The extraction of dates is usually a simple task and can easily be completed by a rule based annotator. If the training data contained more mentions of dates, there is no doubt that the machine learning model would also do better.

Annotating text from medical data requires defining the type system (i.e. the data model). Although universal type systems exist (e.g. UMLS) the type system used for annotation must be adapted to the objective of the annotator. There does not yet exist an annotator capable of annotating all the information contained in a text and there may never be due to opposing interpretations. In the real-world setting, an annotator can only be defined in pair with a defined objective.

The entries from the I2B2 competition obtained results ranging from an F1 score or 0.3 to around 0.9. However, they were trained on the whole set of documents (several hundred) and often used a combination of machine learning and rule based annotators. The results obtained on this small set of training data are encouraging.

4.2 CONCLUSIONS

The first objective of this paper was to demonstrate that a machine learning annotator (produced with Watson Knowledge Studio) can annotate clinical documents and overcome the difficulties inherent to the medical field (complexity of the ontologies, specific formulations etc..). The overall score of 0.71 could certainly be improved by increasing the number of training documents. The second objective was to demonstrate the ease of use of Watson Knowledge Studio. A 2015 paper the Journal of Biomedical Informatics [9] evaluates the ease of adoption of clinical annotation software and concludes that most systems are extremely difficult to use and require a wide set of

skills. Watson Knowledge Studio provides both a high quality machine learning system, and an ergonomic interface and development process that allows the development of high quality annotators.

5 FUTURE WORK

This annotator is a proof of concept (it is possible to annotate clinical documents) and a proof of technology (IBM Watson Knowledge Studio). Many improvements could be made to the annotator. Using a combination of machine learning and rule based annotators could greatly improve the scope of the annotator by including logical rules (e.g. detect if the result of a test is higher or lower than a certain value and annotate if positive). The IBM products Watson Knowledge Studio and Watson Content Analytics can be integrated together to provide a hybrid annotator using both machine learning and rules. The IBM Watson products are evolving quickly and as of January 2017, a new functionality has been added to Watson Knowledge Studio allowing the creation of basic rule based annotations. This functionality provides yet another way to improve the scope and quality of the annotator. The information that could be extracted and used for big data analysis of unstructured clinical reports is limitless. A CAD annotator extracts one type of pathology and could be coupled to other annotators to increase the scope of the metadata added to the clinical reports (Diabetes, Obesity, Medication, etc..).

The aim of this paper was not to describe the technical difficulties linked to information extraction but rather demonstrate how current machine learning technology can be used to produce information extraction systems. It would be interesting to carry out an in-depth study of the technologies that are used by a product like Watson Knowledge Studio to annotate the documents (tokenization, grammatical parsing systems, named entity recognition etc..).

6 REFERENCES

[1] Stephane M Meystre et al., "Automatic de-identification of textual documents in the electronic health record: a review of recent research," 2010.

[2] U.S Department of Health and Human Services, "Health Information Privacy," [Online]. Available: http://www.hhs.gov/hipaa/. [Accessed November 2016].

[3] H. NLP, "ShARe project," [Online]. Available: https://healthnlp.hms.harvard.edu/share/wiki/index.php/Main_Page. [Accessed October 2016].

[4] Informatics for Integrating Biology and the Bedside, "I2B2," [Online]. Available: https://www.i2b2.org/. [Accessed November 2016].

[5] Stubbs A, Uzuner O, "Annotation risk factors for heart disease in clinical narratives for diabetic patients," *J Biomed Inform. 2015 21,* 2015.

[6] "Unstructured Information Management Architecture SDK," IBM, [Online]. Available: https://www.ibm.com/developerworks/data/downloads/uima/index.html.

[7] Laura Chiticariu, Yunyao Li, Frederick R. Reiss, "Rule-based Information Extraction is Dead! Long Live Rule-based Information Extraction Systems!," 2013.

[8] D. Maynard, W. Peters and L. Yaoyong, "Metrics for Evaluation of Ontology-based Information Extraction," 2006.

[9] Kai Zheng et al., "Ease of adoption of clinical natural language processing software: An evaluation of five systems," 2015.

7 APPENDIX: STATE OF THE ART OF CLINICAL NLP

This state of the art provides a broader view of clinical NLP than presented in this papers and should help the reader place the work within its context.

TABLE OF CONTENTS

1 About Natural Language Processing

1.1 The need for NLP

The healthcare industry is becoming a data driven industry. The amount of data now produced is colossal and in the past years the main focus has shifted from trying to produce medical data to trying to analyze it. Analyzing clinical data is a complex task, studies suggest that around 60% of all clinical data is in unstructured format [1]. Unstructured data does not fit a predefined model and is difficult to process, it is usually text heavy and can commonly be analyzed only by humans. In healthcare, examples of unstructured data are the narratives present in electronic health records (EHR) or medical research papers and best practice guides. The information contained in unstructured data is often very informative but because it is not directly usable by computers much of the information is lost. For unstructured medical data to be fully used, it is necessary to design computer systems that can understand text in unstructured form. Natural language processing (NLP) is the study of interactions between the computer and human (natural) languages [2].

NLP technologies will ultimately allow computers to read through huge amounts of written unstructured data and produce output. This state of the art will focus on a subfield of NLP called natural language understanding (NLU) which is currently the focus of NLP research in healthcare. NLU is the study of how computers understand human languages [3]. To remain consistent with the literature, NLP will usually refer to clinical natural language understanding which is the use of natural language understanding in healthcare.

1.2 NLP is complex

NLP is complex. Before trying to produce a computer system that understands, it is necessary to reflect on what it means to understand. Understanding a statement is far more complex than understanding the individual words that compose it. NLP systems divide understanding into layers of knowledge. The layers go from the basic blocks that compose words to the global understanding of a sentence in its context [4]. In text analysis, the most basic level is the morphologic knowledge, which refers to how words are built with basic units (morphemes). For instance, *smaller* is composed of *small* and *er* which together indicate a comparison of the size of two entities. Lexical knowledge refers to the meaning of individual words, a same word can have several potential meanings. Syntactic knowledge refers to how words are put together to form a sentence. Semantic knowledge is the understanding of how words that are put together form a sentence with a meaning. Discourse knowledge refers to the understanding of a sentence within a larger group of sentences, for instance the study of pronouns referring to an object previously mentioned. Finally, pragmatic knowledge describes how world knowledge or knowledge of facts can contribute to the meaning of a sentence.

All these levels of understanding are natural for humans but for a machine to fully understand, it must somehow integrate knowledge of all these levels. This complexity allows for ambiguities in human language. Moreover, the medical field uses many acronyms, some acronyms can mean different things depending on the context. For instance, APC can mean Activated Protein C or

AlloPhyCocyanin amongst other things [5]. Other aspects that add complexity are different languages, spelling errors, typos etc. [6]

1.3 THE PRIMARY USES OF NLP

Following are some NLP use cases [5]. Many are based on information extraction which is the analysis of natural language to extract data or meaning [7]. The Open Health Natural Language Processing Consortium (OHNLP) defines several NLP use cases in the clinical domain: [8]

- Patient cohort identification: The aim of patient cohort identification is to group patients depending on pre-defined characteristics. Today the cohort identification mechanisms are extremely costly in time and resources. It is now possible to define inclusion or exclusion criteria for the cohort identification and search structured and unstructured data simultaneously.
- Clinical decision support: developing systems that can aid clinical experts in their decision making. Clinical decision support usually relies on the analysis of large amounts of data that could not be read by a human and suggests a course of action or give information that could help the clinician in his decision making.
- Health care quality research: analysis of physician's observations on patients to determine the quality of healthcare given to patients.
- Personalized medicine: patient medication history is of extreme importance in defining how a patient reacts to certain medication and how effective a treatment is for a given patient. Often however, most of the information on the effectiveness of medication is written in unstructured form in clinical narratives. Being able to analyze this unstructured data could help personalize treatment.
- Drug development: NLP can help determine drugs that could be repurposed based on large EHR sets. It could also be possible to analyze large sets of written reports on patients' reactions to medication to discover adverse drug effects.
- Text summarization: There are two cases in which text summarization can be used. The first one is to summarize data on patients that may be present over several clinical documents. This would help clinicians detect the key information on a patient without having to go through the whole medical record. The second use is in the summary of clinical texts. Clinicians could use text summarization to summarize large sets of clinical data such as best practices or research papers to determine the key elements of these documents.

1.4 THE BASIC COMPONENTS OF AN NLP SYSTEM

NLP systems use a wide variety of technologies to achieve language understanding, following are some of the major tasks that almost all NLP systems use [9]. Tokenization or word segmentation is the splitting of spans of text into tokens. In English, tokens are usually individual words so tokenization is a fairly easy task. However, in other languages such as Chinese, tokenization can be challenging. Part of speech tagging (POS) is the tagging of individual words to syntactic functions. In English, this means determining whether a word is a noun, verb, adjective, determiner

etc. Named entity recognition is an extremely important aspect of NLP, it is the task of dividing words into pre-defined categories (often called entities). These categories can represent many things, events, dates, locations, people, organizations etc. In Clinical NLP, entities can be body parts, medical conditions etc. Another aspect of NLP is relationship extraction. Relationship extraction aims at determining relationships between entities found in the Named Entity Recognition task. An example of a relationship between a body part entity and a medical injury entity could be: "injury located at body part". Other important aspects include negation detection, which is especially important in healthcare. The difference between: "The patient had no symptoms" and "the patient had symptoms" is extremely important and NLP tools must be able to detect negation as well as nuance (i.e. severe injury or minor injury). Synonym extraction and abbreviation expansion are other tasks [10]. These are only some of the tasks required for NLP and depending on the goal and complexity of the application, a NLP system may integrate more or less of these components.

2 REVIEW OF NLP SOFTWARE AND NLP INITIATIVES

This section aims to provide an overview of the current clinical NLP systems and the community initiatives that led to the development of the systems used today. A lot of the clinical NLP systems are open source and active communities are at work to produce high quality clinical natural language processing tools. The following list will not include the Watson software because Watson will be the main focus of the rest of the thesis.

2.1 CLINICAL NLP ACTORS

Clinical NLP being such an important field in healthcare, many actors are present. Following are a few of the actors and communities that play an interesting role, the list is by no means exhaustive.

Today, many clinical NLP tools rely on the UIMA (Unstructured Information Management Applications) software systems. UIM applications are software systems that analyze unstructured information (including text) to produce relevant information. IBM produced the UIMA pipeline and has since donated the source code to the Apache Software Foundation, the source code is now open to NLP communities. The Apache Software Foundation is an open source community of developers. The UIMA project is under the Apache V2.0 license that does not restrict the development of applications, therefore anyone can build their own NLP software on top of the main pipeline [11].

Another important actor in the field of clinical NLP is the Open Health Natural Language Processing Consortium (OHNLP). The following is extracted from their website and summarizes their goal: "The goal of the Open Health Natural Language Processing Consortium is to establish an **open source** consortium to promote past and current development efforts and to encourage participation in advancing future efforts. The purpose of this consortium is to facilitate and encourage new annotator and pipeline development, exchange insights and collaborate on novel biomedical natural language processing systems and develop gold-standard corpora for development and testing. The Consortium promotes the open source UIMA framework and SDK as the basis for biomedical NLP systems." [8]. The OHNLP consortium was founded in 2009 by IBM and the Mayo clinic, both parties at the time contributed software to the project. IBM contributed a software called MEDKAT/p which is trained to extract cancer specific characteristics from unstructured text. The Mayo Clinic contributed several software tools called MedTagger for information extraction based on patterns and named entity recognition, MedTime for the detection of temporal expressions from clinical text and MedXN for extraction and normalization, it can for instance provide abbreviation expansion [12]. Today, other contributors have joined the consortium and provided their own set of clinical NLP software. The OHNLP provides "best of breed" software that can be integrated into more global systems. They encourage the use of the Apache V2.0 license although they also support tools developed under other licenses such as GNU General Public License. Another aim of the OHNLP is to provide guidelines for better interoperability between the clinical NLP software tools.

A very interesting resource to scope the clinical NLP software is the Online Registry for Biomedical Informatics Tools (ORBIT). Their role is summarized in this extract taken from their website [13]: "The Online Registry of Biomedical Informatics Tools (ORBIT) Project is the result of a collaboration of more than 30 researchers, developers, informaticians, etc. across more than a dozen academic and federal research organizations. It was created to provide researchers and developers with a single, simple way to register and find open source software resources. ORBIT was created by and for biomedical informatics researchers, leading to features we hope will be useful for others in the community." Orbit is a registry for all open source clinical informatics tools, however it is possible to search by field. A simple search for NLP tools in healthcare comes up with 70 results. This number shows how important NLP is for healthcare. In the following part, we will present some of the most used and well known clinical NLP tools.

NLP Actors	Role
IBM	Creating the UIMA framework
Apache	Maintain and improve UIMA framework
OHNLP (Open Health NLP)	Consortium to promote the development of open clinical NLP tools
ORBIT (Online Registry for Biomedical Informatics Tools)	Provide an online registry for open clinical NLP tools

Table 4 Actors in clinical NLP

2.2 NLP SOFTWARE

Before discussing any of the software tools used for clinical NLP it is essential to define the Unified Medical Language System (UMLS) [14] [15]. The UMLS is a set of resources that group biomedical vocabularies and standards to facilitate interoperability between clinical NLP systems. There are three UMLS tools: Metathesaurus, which contains terms and codes from many medical vocabularies; Semantic Network, which defines broad categories and their relationships; and a set of natural language processing tools. The Metathesaurus is the largest element of the UMLS. It groups an enormous set of biomedical terms and organizes them by concept and meaning. The UMLS is used by many clinical NLP tools to provide a standard for relationships and terms that improve interoperability between the NLP modules. It is possible to download the Metathesaurus for free and many resources are available to help use it and integrate it into the software.

With the development of the UMLS concepts, came the development of a tool called MetaMap [16] [17]. Extracted from their website is the following summary: "**MetaMap** is a highly configurable program developed by Dr. Alan (Lan) Aronson at the National Library of Medicine (NLM) to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text. MetaMap uses a knowledge-intensive approach based on symbolic, natural-language processing (NLP) and computational-linguistic techniques. Besides being applied for both IR (information retrieval) and data-mining applications, MetaMap is one of the foundations of NLM's (National Library of Medicine) Medical Text Indexer (MTI) which is being used for both semiautomatic and fully automatic indexing of biomedical literature at NLM."

A very prominent clinical NLP tool is cTakes [18]. It was first developed in 2006 by a group of software engineers, physicians and engineers at the Mayo Clinic. Now an open source project under the Apache 2.0 license, it is currently used by the Mayo clinic and integrated to their system. It has processed more than 80 million documents [19]. Ctakes' goal is to be a world class clinical NLP system integrating the best of breed modules from various sources. For instance, the OHNLP has contributed NLP modules to the cTakes project [20]. The medTagger developed by the Mayo clinic is part of the cTakes software. CTakes is UIMA based and provides NLP modules such as tokenization, dependency parsing, semantic processing, part of speech tagging or named entity recognition [21] [22]. The cTakes software includes several modules such as UIMA CVD (UIMA CAS Visual Debugger). It provides tagging capabilities on raw text. UIMA CPE is also present and provides the ability to process multiple documents in a batch. Other modules include the cTakes Graphical user interface (GUI) and TimeLanes to extract temporal information.

Clinical NLP software and tools	**Main target**
UMLS	Provide standardized ontology for clinical NLP systems
MetaMap	Clinical text mining and information retrieval
cTakes	Clinical text mining
UIMA	Framework for developing NLP software

Table 5 Clinical NLP software

3 EXAMPLES OF CLINICAL IMPLEMENTATIONS OF NLP

This section will present some concrete applications of clinical NLP. A first application was presented by Rajakrishnan Vijayakrishnana et al. in a 2014 paper called "Prevalence of Heart Failure Signs and Symptoms in a Large Primary Care Population Identified Through the Use of Text and Data Mining of the Electronic Health Record" [23]. This paper suggests that electronic health records (EHR) contain tremendous amounts of data that, if correctly analyzed, could lead to earlier detection of heart failure. They carried out a study of 50,000 primary care patients to identify signs and symptoms of heart failure in the years preceding the diagnosis. They found a total of 892,805 criteria over the 50,000 patients. Of the 4644 incident HF cases, 85% had at least one criterion within a year period prior to their HF. This study shows a concrete case in which NLP analysis of EHR could provide signs of HF that would otherwise have been missed.

A second application is the use of clinical NLP to detect binge eating disorder (BED) in patients from an analysis of their EHR [24]. The study used EHRs from the Department of Veterans Affairs between 2000 and 2011. NLP methods identified 1487 BED patients which corresponds to an accuracy of 91.8% and a sensitivity of 96.2% compared to human review.

A third application is again based on the analysis of EHR from patient to assess prostate biopsy results [25]. The aim of the study was to determine patients with prostate cancer from their EHR. The first part of the study was to read pathology reports and extract variables that were representative of prostate cancer. These variables were then searched by an internal NLP program in a series of 18.453 pathology reports. The results were very encouraging, the NLP correctly detected 117 out of 118 patients with prostatic adenocarcinoma.

These studies demonstrate the possibilities for clinical NLP. Although they only represent a small scope of what can be done in NLP, they are indicative of the potential of clinical NLP.

4 DIFFICULTIES AND PROBLEMS IN CURRENT NLP SYSTEMS

This section will present some of the difficulties in the NLP field that have impeded progress in the past or still pose a problem today.

Many of the NLP technologies are based on Machine Learning technologies. By definition these technologies rely on large amounts of data to train and improve results. However, most of the clinical training data contains Protected Health Information (PHI). This information means that it is extremely difficult to distribute sets of training data even for research purposes. For a long time, this slowed the advances in clinical NLP. In 2010, Meyster et al. [26] published a review of the de-identification methods. Since then many studies have been done and the software is sufficiently effective. They use mostly machine learning algorithms to detect the presence of personal data. De-identification is not as simple as just removing names, any data that could help identifying the person must be removed. The de-identification techniques are NLP based. There are now many corpora of de-identified medical documents that can be used for training clinical NLP data. The Shared Annotated Resources (ShARe) project provides clinical data that helps train clinical NLP systems [27].

An interesting paper published in 2015 in the Journal of Biomedical Informatics [28] evaluates the ease of adoption of some of the NLP software tools. The clinical NLP tools tested in the study were BioMEDICUS, CliNER, MedEx, MedXN, MIST. These tools are representative of the open source software that is available to the public for clinical NLP and target different aspects of NLP (named entity recognition, information extraction etc...). The study evaluated the ease of adoption by asking people to download, install, and use the product. They also asked them to rate how easy it was to understand the objective of the NLP tool and the expected output of the annotation. This is very important because NLP is a complex field and the tools provided cover a wide range of applications from de-identification to text summary. Moreover, most tools only actually cover the low level processing tasks and need to be implemented with other tools to provide high level capabilities. Without understanding what a specific piece of software does, it is difficult to put together a pipeline that will perform a specific task. The study found that for non-savvy users it was very difficult to understand both what the software did and how to use it.

5 SHORT SUMMARY ON IBM WATSON AND USE IN NLP

IBM Watson is a Natural Language Processing system (NLP). It works on many aspects of NLP and has an extremely broad range of applications. IBM Watson's NLP technologies can be used to drive forward the use of NLP in healthcare. IBM's objective is to provide a simple way for non-technical users to create annotation tools for any type of text data including clinical texts.

Despite being a very active area of research, clinical NLP is not yet widely spread in the healthcare sector. There are many reasons for this, including the high technicality of many of the tools and the high specificity of NLP applications, even within healthcare make it difficult to provide a tool that fits all needs. Through Watson, IBM is providing NLP tools and services to develop applications for specific use cases. Watson For Oncology has already broken ground in clinical NLP. It uses NLP to extract information from clinical research papers and guidelines to suggest cancer treatment paths depending on the patient information.

NLP systems can either be rule based or statistical. Historically, most NLP systems were rule based. With the development of machine learning and the increased availability of annotated data, statistical algorithms are become more accessible and popular. IBM Watson uses both rule based and statistical methods to provide a wide range of capabilities in NLP.

6 REFERENCES

[1] DataMark Incorporated, "Unstructured Data in Electronic Health Record (EHR) Systems: Challenges and SolutionsHealthcare," 2013.

[2] Chowdhury, Gobina G, "Natural Language Processing,", 2003.

[3] T. Winograd, "Understanding natural language," 2004.

[4] R. K. M. Ronilda Lacson PhD, "Natural Language Processing: The Basics," *Journal of American College of Radiology,* 2011.

[5] O. G. Iroju and O. J. Olaleke, "A Systematic Review of Natural Language Processing in Healthcare," 2015.

[6] A. Névéol, P. Zweigenbaum, Section Editors for the IMIA Yearbook Section on Clinical, "Clinical Natural Language Processing in 2014: Foundational Methods Supporting Efficient," 2015.

[7] A. Moschitti, "Natural Language Processing and Automated Text Categorization," 2003.

[8] O. H. N. L. P. Consortium, "OHNLP Main Page," [Online]. Available: http://www.ohnlp.org/index.php/Main_Page. [Accessed October 2016].

[9] Stephen T. Wu, Vinod C. Kaggal, Guergana K. Savova, Hongfang Liu, Dmitriy Dligach, "Generality and Reuse in a Common Type System for Clinical Natural Language Processing".

[10] Aron Henriksson, Hans Moen, Maria Skeppstedt, Vidas Daudaravicius and Martin Duneld, "Synonym extraction and abbreviation expansion with ensembles of semantic spaces," 2014.

[11] "Unstructured Information Management Architecture SDK," IBM, [Online]. Available: https://www.ibm.com/developerworks/data/downloads/uima/index.html.

[12] O. H. N. L. P. Consortium, "Tool list," [Online]. Available: http://www.ohnlp.org/index.php/OHNLP_Tool_List. [Accessed September 2016].

[13] Tools, Online Registry for Biomedical Informatics, "About Orbit," [Online]. Available: https://orbit.nlm.nih.gov/about. [Accessed September 2016].

[14] NLM, "Unified Medical Language System," [Online]. Available: https://www.nlm.nih.gov/research/umls. [Accessed October 2016].

[15] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," 2003.

[16] P. Alan R. Aronson, "Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program," 2001.

[17] MetaMap, "MetaMap - A Tool For Recognizing UMLS Concepts in Text," [Online]. Available: https://metamap.nlm.nih.gov/.

[18] Guergana K Savova et al., "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," 2010.

[19] cTakes, "History of cTakes," [Online]. Available: http://ctakes.apache.org/history.html. [Accessed September 2016].

[20] Masanz, James ; Pakhomov, Serguei V ; Xu, Hua ; Wu, Stephen T ; Chute, Christopher G ; Liu, Hongfang, "Open Source Clinical NLP – More than Any Single System," 2014.

[21] S. Velupillai, D. Mowery, B. R. South, M. Kvist, H. Dalianis, "Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis," 2015.

[22] Guergana K Savova et al., "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," 2010.

[23] Rajakrishnan Vijayakrishnan et al., "Prevalence of Heart Failure Signs and Symptoms in a Large Primary Care Population Identified Through the Use of Text and Data Mining of the Electronic Health Record," 2014.

[24] Brandon K Bellows et al., "Automated identification of patients with a diagnosis of binge eating disorder from narrative electronic health records," 2013.

[25] Anil A. Thomas et al., "Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results," 2013.

[26] Stephane M Meystre et al., "Automatic de-identification of textual documents in the electronic health record: a review of recent research," 2010.

[27] H. NLP, "ShARe project," [Online]. Available: https://healthnlp.hms.harvard.edu/share/wiki/index.php/Main_Page. [Accessed October 2016].

[28] Kai Zheng et al., "Ease of adoption of clinical natural language processing software: An evaluation of five systems," 2015.

TRITA 2017:3