Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By ANAND KRISHNAN

Entitled

MINING CAUSAL ASSOCIATIONS FROM GERIATRIC LITERATURE

For the degree of MASTER OF SCIENCE

Is approved by the final examining committee:

Dr. MATHEW J. PALAKAL
Chair

Dr. YUNI XIA

Dr. ARJAN DURRESI

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): MATHEW J. PALAKAL

Approved by: SHIAOFEN FANG 7/2/2012
Head of the Graduate Program Date

Graduate School Form 20
(Revised 9/10)

PURDUE UNIVERSITY
GRADUATE SCHOOL

Research Integrity and Copyright Disclaimer

Title of Thesis/Dissertation:
MINING CAUSAL ASSOCIATIONS FROM GERIATRIC LITERATURE

For the degree of MASTER OF SCIENCE

I certify that in the preparation of this thesis, I have observed the provisions of *Purdue University Executive Memorandum No. C-22,* September 6, 1991, *Policy on Integrity in Research.**

Further, I certify that this work is free of plagiarism and all materials appearing in this thesis/dissertation have been properly quoted and attributed.

I certify that all copyrighted material incorporated into this thesis/dissertation is in compliance with the United States' copyright law and that I have received written permission from the copyright owners for my use of their work, which is beyond the scope of the law. I agree to indemnify and save harmless Purdue University from any and all claims that may be asserted or that may arise from any copyright violation.

ANAND KRISHNAN
_____
Printed Name and Signature of Candidate

7/2/2012
_____
Date (month/day/year)

*Located at http://www.purdue.edu/policies/pages/teach_res_outreach/c_22.html

MINING CAUSAL ASSOCIATIONS FROM GERIATRIC LITERATURE

A Thesis

Submitted to the Faculty

of

Purdue University

by

Anand Krishnan

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

August 2012

Purdue University

Indianapolis, Indiana

This work is dedicated to my family.

ACKNOWLEDGMENTS

I am heartily thankful to my supervisor, Dr. Mathew J. Palakal, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject.

I want to thank Dr. Yuni Xia and Dr. Arjan Durresi for agreeing to be a part of my Thesis Committee.

I also want to thank Jon Sligh, Natalie Crohn, Heather Bush, Eric Tinsley and Jason De Pasquale from Alligent and Jean Bandos for their valuable support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Krishnan, Anand. M.S., Purdue University, August 2012. Mining Causal Associations from Geriatric Literature. Major Professor: Mathew J. Palakal.

Literature pertaining to geriatric care contains rich information regarding the best practices related to geriatric health care issues. The publication domain of geriatric care is small as compared to other health related areas, however, there are over a million articles pertaining to different cases and case interventions capturing best practice outcomes. If the data found in these articles could be harvested and processed effectively, such knowledge could then be translated from research to practice in a quicker and more efficient manner. Geriatric literature contains multiple domains or practice areas and within these domains is a wealth of information such as interventions, information on care for elderly, case studies, and real life scenarios. These articles are comprised of a variety of causal relationships such as the relationship between interventions and disorders. The goal of this study is to identify these causal relations from published abstracts. Natural language processing and statistical methods were adopted to identify and extract these causal relations. Using the developed methods, causal relations were extracted with precision of 79.54%, recall of 81% while only having a false positive rate 8%.

1 INTRODUCTION

1.1 Overview

Modern day science has an abundance of data. This data can be derived from various different sources like public databases, repositories, collaborations, etc. Yet the more useful knowledge remains trapped in the literature. Computational methods have evolved to handle large amounts of text and derive knowledge from it. This applies to the field of geriatrics as well. Text mining enables analysis of large collections of unstructured or semi-structured documents for the purposes of extracting interesting and non-trivial patterns or knowledge [1].

The field of geriatrics presents wealth of information that is derived from studies conducted in multitude of locations, such as nursing homes and hospitals. Geriatric literature is comprised of documents that contain information about Geriatric Syndromes [2]. These syndromes are groups of specific signals and symptoms that occur more often in the elderly and can impact patient morbidity and mortality. Normal aging changes, multiple co-morbidities, and adverse effects of therapeutic interventions contribute to the development of Geriatric Syndromes. These syndromes are becoming increasingly important for nurses and care providers to consider as the patient population ages. In fact this development has been included in AACNs 2006 edition of its Core Curriculum for Critical Care Nursing. It has been reported that on an average, 35% to 45% of people above the age of 65 experience a fall annually. Studies have also shown that there are 1.5 falls per bed amongst the people of age 65 and above. Numerous publications are available regarding the best practices for geriatric care to address Geriatric Syndromes and other geriatric related issues. Though the number of publications specific to geriatric care is small, there are millions of pub-

lished peer-reviewed articles that contain different interventions, use-case scenarios, and problems that the elderly face. There is no standard corpus for all these cases and interventions, and there is no significant work done in this area. Mining this kind of literature can be extremely challenging as the data is scattered over multiple domains. One way of collecting data is to capture the abstracts that provide a synopsis of what the article contains and apply mining techniques like Pattern Recognition, Classification, Neural Networks, Support Vector Machines, and Cluster Analysis to extract relevant information from them [3] [4] [5] [6] [7] [8]. In this paper a multi-layered model is applied to extract relevant information in the form of causal associations from the abstracts. The goal of model is to clarify complicated mechanisms of decision-making processes and to automate these functions using computers [9].

1.2 Information Extraction from Literature

Typically a text mining system begins with collections of raw documents that does not contain any annotations, labels or tags. These documents are then tagged automatically by categories, terms or relationships that are extracted directly from the documents. The extracted categories, terms, entities and relationships are used to support a range of data mining operations on the documents [10]. Figure 1.1 shows the typical Information extraction process.

The task of Information Extraction (IE) systems is extracting structured information from unstructured documents. Several IE systems have been developed to help researchers extract, convert and organize new information automatically from textual literature. These are employed majorly to draw out relevant information from biological documents like extracting protein and genomic sequence data.

1.3 Geriatric Literature

Geriatric literature contains rich information regarding the “best practices” related to geriatric health care issues. There are over a million articles that bear

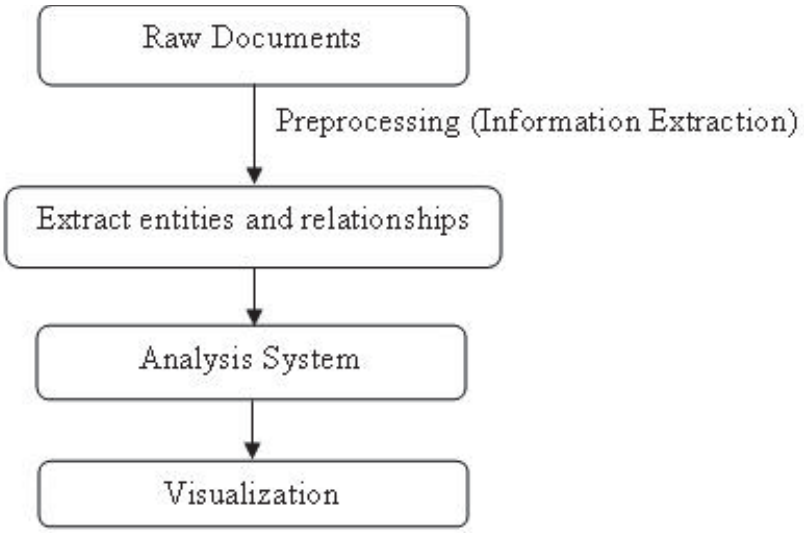

Figure 1.1.: Text Mining Process

information about various "case" and case "interventions" (cause and effect) data. This can be processed and translated from using an Information Extraction system in a quicker and more efficient manner.

The field of Geriatrics requires expertise that only a few individuals possess. These individuals are referred to as domain experts. After initial analysis for this project, the domain experts chose 42 of the most common Geriatric Syndromes. Table 1.1 shows the list of all Care Categories identified for this study.

1.4 Goal of the Research

The goal of this thesis is to extract causal relations from geriatric abstracts and process it further to build a knowledgebase of geriatric care information that can be used by care providers. The system would identify causal relations which would fit into a Bayesian model as part of a decision support system. The model identifies such sentences and classifies them into two classes; Causal and Non-Causal.

Table 1.1: Care Categories

Fall Risk	Financial	Care Provision
Cognition	Nutrition	Health History
Incontinence	Instrumental Activities Of Daily Living (IADLS)	Social
Wellness Prevention	Mobility	Well-Being
Depressive Symptoms	Safety	Supportive Services
Health Status	Providers	Safety and Assistive Devices
Caregiver Support	Anxiety	Elder Abuse
Pain Management	Environmental	Information Preference
Legal	Emotional	Intellectual
Sensory	Medical Issues	Social Interaction
Substance Abuse	Insurance Issues	Preferences
Stress Management	Medication Management	Legal Older Adults
Alternative Living Options	Activities Of Daily Living	Medical Alerts
Sleep	Spiritual	Chronic Disease

1.5 Contribution of the Thesis

The proposed system in this thesis uses a new technique of integrating Syntactic tagging, Semantic tagging, Dictionaries and Conditional Random Fields for extraction of causal relations from Geriatric abstracts. This is a stand-alone system that would be the engine to provide quality information in the form of causal relations to a decision support system.

The system will have information extracted from a collection of 2280 Pubmed [11] abstracts pertaining to the field of geriatric care. The results produced by this framework will enhance the of information extraction systems in identifying quality causal sentences and even predict new actors that may appear in future articles.

2 RELATED WORK

Information Extraction dates back to the late 1970s. A significant amount of research has been done in the area of information extraction from literature. There are different types of relationships that can be extracted from literature and there are several methods that have been used to obtain this information. These methods can be broadly classified into deterministic or probabilistic based methods. Deterministic methods are not very scalable to new domains while probabilistic methods are more flexible in their implementation. The relation extraction can also depend on the type of domain that is under study. Causal relations can be expressed in different ways and they can differ from domain to domain. It can be expressed between two sentences, between two phrases, between subject and object noun phrases, in intra-structure of noun phrases and even between paragraphs that describe events. Some methods make use of a combination of deterministic and probabilistic approach for information extraction. This chapter describes the work done in information extraction using deterministic and probabilistic methods.

2.1 Natural Language Processing

Natural Language Processing (NLP) is an area of research that explores how natural language text can be understood and manipulated by computers to do useful things [12]. [13] states it as a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts. The purpose of this computation is to achieve human-like language processing for a range of tasks or applications. For any effective information extraction, techniques derived from natural language processing are used. A graphical representation of NLP in Figure 2.1 shows the most important components of a NLP process. These components are

implemented in a number of ways using a combination of approaches - deterministic, probabilistic, automatic, semi-automatic, rule-based etc. to extract the required knowledge.

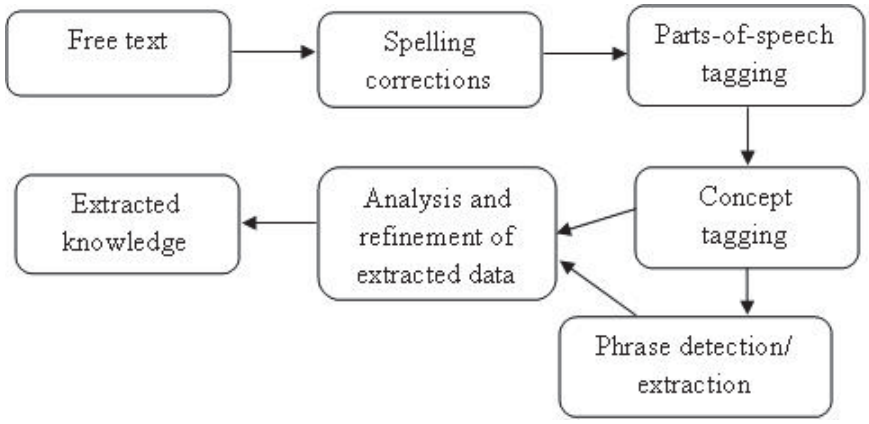

Figure 2.1.: Overview of NLP Process

2.1.1 Syntactic Tags - Parts-Of-Speech Tagging POS

For natural language, syntax provides rules or standardized features to put together words to form components of sentence. Syntactic features describe how a certain token relates to others. In other words, an indication is given of the functional role of the token. The process of Parts-Of-Speech tagging is to identify a contextually proper morpho-syntactic description for each ambiguous word in a text. [14].

A major aspect of Natural language processing is the Parts-of-Speech tagging. Natural language has several different parts of speech that include nouns, pronouns, verbs, adjectives, adverbs, prepositions, conjunctions and interjections. When a sentence is passed through a tagging process, the natural language text is assigned its parts of speech. There are several other POS tagging tools such as Brill Tagger [15] which has an accuracy of 93-95%. The Stanford POS tagger [16] provides an accuracy of upto 97%. The Medpost [17] POS tagger has an accuracy of 97% which is one of the most popular tagging tools. Example for Medpost POS Tagging.

Fall rates among community-dwelling elderly people increase with age and are greater for women than men.

Figure 2.2.: Sentence Before Medpost POS Tagging

Fall_JJ rates_NNS among_II community-dwelling_VVNJ elderly_JJ people_NNS increase_VVB with_II age_NN and_CC are_VBB greater_JJR for_II women_NNS than_CSN men_NNS ._.

Figure 2.3.: Sentence After Medpost POS Tagging

Figure 2.3 shows the POS tagged output of Medpost Tagger of the sentence shown in Figure 2.2. The tags suffixed to each word are used by various NLP tools.

2.1.2 Extracting Causal Associations

Sentences like *"Inflation affects the buying power of the dollar."*, *"Cigarette smoking causes cancer."*, *"Happiness increases with sharing."*, *"Guitar is an instrument associated with music."* very clearly shows a relation between one event or entity (Inflation, Cigarette, Happiness, Guitar) to another entity (buying power, cancer, sharing, music) with the help of temporal relations like *"affect"*, *"causes"*, *"increases"* and *"associated"*. Examples such as these that are used in common language are indicative of the ubiquity of causality in everyday life. One or the other ways, causality affects us all as it expresses the dynamics of a system. Extraction of such causal relations from any literature can be very tricky if we understand the complex nature of natural language.

Early research in causal association extraction analysis started with a manually curated causal pattern set to find causal relationships from literature. The literature under study was run through these set of patterns and the required information was extracted.

The causal patterns Khoo et al. [18] investigated an effective cause-effect information extraction system from newspaper using simple computational method. They demonstrated an automatic method for identifying and extracting cause-effect infrmation in text from the Wall Street Journal using linguistic clues and pattern-matching. They constructed a set of linguistic patterns after a thorough review of the literature and on sample Wall Street Journal sentences. The results obtained from this method were verified by two human experts. The linguistic patterns developed in the study were able to extract about 68% of the causal relations that are clearly expressed within a sentence or between adjacent sentences. The study also reported some errors by the computer program that was caused mainly due to complex sentence structures, lexical ambiguity and an absence of inference from world knowledge. This method provided a deterministic approach which shows that causal extraction can be achieved if the linguistic patterns collected from the literature have a wider coverage and is generalized to work for any domain. Techniques have been developed using inter-sentence lexical pair probability for differentiating the relations between sentences. Marcu et al. [19] hypothesized that lexical item pairs can help in finding discourse relations that hold between the text spans in which the lexical items occur. In their study they used sentence pairs connected with the phrases because and thus to distinguish the causal relation from other relations. There were two problems to test this hypothesis. The first was to acquire knowledge about CONTRAST relations, for example, word-pairs like good-fails and embargo-legally indicate contrast relations. They built a table that contains contrasting word-pairs to address this problem. The second problem was to find a means to learn which pairs of lexical items are likely to co-occur with each disclosure relation and how to apply the learned information on any pair of text spans and to determine disclosure relation between them. They used a Bayesian probabilistic framework to resolve this problem. This method used only nouns, verbs and cue phrases in each sentence/clause. Non-causal lexical pairs were also collected from the sentence pairs to compose the Naive Bayes classifier. The result shows an accuracy of 57% in inter-sentence causality extraction. From

this, it can be understood that lexical pair probability contributes to the causality extraction. Since this work involved extraction of phrases that connect the sentence pairs, causality extraction problem can be addressed by building a dictionary of such causal words extracted from literature.

Causal relation extraction can also be done in a semi-automatic form. The method presented by [20] shows one such semi-automatic method of discovering generally applicable lexico-syntactic patterns that refer to the causal relation. The patterns are discovered automatically, but their validation is done semi-automatically. They discuss several ways in which a causal relation can be expressed but focus on a single form, *<NounPhrase1 verb NounPhrase2>*. Lexico-syntactic pattern are discovered from a semantic relation for a list of noun-phrases extracted from Wordnet 1.7 [21] and patterns are extracted that links the two selected noun phrases by searching a collection of texts. This gave a list of verb/verbal expressions that refer to causation. Once the list is formed, the noun phrases in the relationship of the form *<NounPhrase1 verb NounPhrase2>* can express explicit or implicit states. Only certain types of such states were considered for the study. These relationships are analyzed and ranked. The result obtained for this experiment used the TREC-9 (TREC-9 2000) collection of texts which contains 3GB of news articles from Wall Street Journal, Financial Times, Financial Report, etc. The results were validated with human annotation. The accuracy obtained by the system in comparison with the average of two human annotations was 65.6%.

2.1.3 Semantic Tagging

Semantic tagging is a method of assigning tags, symbols or markers to text strings which can help in identifying their meaning so that the string and its meaning can be made discoverable and readable not only by humans but also by computers. It involves annotating a corpus with instructions that specifies various features and qualities of meaning in the corpus [22]. There are several systems in which semantic tagging is

being applied. In each of these systems, the words in the corpus are annotated with various strategies referring to their meanings and these strategies can vary from one domain to another. The simplest example of such a tagging scheme is the parts-of-speech tagger where in the where it assigns a grammatical category (noun, verb, pronoun, etc.) to each token in the text. Another example of such tagging scheme can be seen in the field of human anatomy. Here we can semantically tag the various parts of body into different categories like eyes can be given the tag Part of Face and heart can be tagged as Internal Organ.

The study in [23] shows the implementation of Sense Tagging which is a process of assigning a particular sense from some vocabulary to the content work in a text. This study discusses the approaches that are applied for Word Sense Disambiguation (WSD). Word sense disambiguation is an open problem in NLP. It provides rules for the identification of the sense of a word in a sentence. The most famous example is *"Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy."* Here, the word pen has at least 5 different meanings and it is a difficult task for a computer system to predict the right sense of the word. Studies have been done on building WSD systems that can achieve consistent accuracy levels in pointing out and possibly, identifying the right word to fix the problem. Sense tagging is very useful since the tags that are added during sense tagging have abundant knowledge and are likely to be extremely useful for further processing. The method discussed here implemented the tagger in three modules.

- Dictionary look-up module: Here the system would stems the words leaving out the sentences and the roots. The stop words are removed and with the help of a machine readable Longman Dictionary for Contemporary English (LDOCE), the meaning of each of the remaining word is extracted and stored.

- Parts-of-speech filter: This step involved tagging the text using Brill Tagger [24] and a translating the text using a defined mapping from syntactic tags assigned by Brill to a simple part-of-speech category that is associated with the LDOCE. All the inconsistent senses are then removed assuming that the tagger has made an error.

- Simulated annealing: In the final stage, an annealing algorithm is used to optimize the dictionary definition overlap for the remaining sentence. At the end of this algorithm, a single sense is assigned to each token which is the tag associated with that token.

This work shows that semantic tagging can be used efficiently on text so improve the understandability of the text by adding more features to them and easing the further processing of the text with other methods.

The tests of this approach were performed on 10 hand-disambiguated sentences from the Wall-Street Journal. Though the test set was small, the performance the tagger was found to be 86% for words which had more than 1 homograph and 57% of tokens were assigned the correct sense using the simple tagger.

The research work performed by [25] talks about detecting signals (presence of data modules) in textual material. This approach makes use of Semantic Tagging method to regulatory signal detection to enhance existing text mining methods. The technical challenges that hamper achieving effective signal detection include:

- Mining unstructured data,

- Increasing document collections, and

- Presence of multi-domain vocabulary.

Lack of annotation and multi-domain vocabulary makes traditional mining techniques ineffective. There are several ways to approach the problem of signal detection.

- A typical idea of using a dictionary or bag-of-words text mining can be used to detect actors in textual material. This approach is not scalable if any new actors were to be added to the domain which would make it a very inefficient approach.

- A semantic text mining framework using information retrieval and extraction techniques for signal detection has also been developed to resolve this problem.

- A learning model that can be trained with several samples of sentences with actors. This is a more scalable and efficient technique since it does not work on a finite set of list or rules.

2.1.4 Conditional Random Field

Assigning label sequences to text is a common problem in many fields, including computational linguistics, bioinformatics and speech recognition [26] [27] [28]. The most common task in NLP is labeling the words in a sentence with its corresponding part-ofspeech tag. There are other kinds of label sequences. For example, labeling cause and effect terms in a sentence or even labeling places, people or organizations in sentences that can be identified for machine learning. The most commonly used method used is employing hidden Markov models [29]. HMMs are a form of generative model, that defines a joint probability distribution *p(x,y)* where x and y are random variables respectively ranging over observation sequences and their corresponding label sequences. In order to define a joint distribution of this nature, generative models must enumerate all possible observation sequences a task which, for most domains, is intractable unless observation elements are represented as isolated units, independent from the other elements in an observation sequence. The means that the observation element at any given instant in time may only directly depend on the state, or label, at that time. Although this assumption can be made for simple data

sets, most real-world can be represented the best if represented in terms a multiple interacting features over a long-range dependency between observation elements.

CRFs are undirected graphical models that model the conditional distribution *p(x|y)* rather than joint probability distribution *p(y,x)* and trained to maximize the conditional probability of outputs given the inputs [30]. The main advantage of CRF over hidden Markov model being its conditional nature which helps in relaxing the independence assumptions required by HMMs in order to ensure tractable inference. Also, CRFs avoid the label bias problem, which is a weakness shown by Maximum Entropy Markov Models (MEMMs) and other conditional Markov models based on directed graphical models. CRF surpasses the performance of both MEMMs and HMMs on a number of real-world tasks.

A probability distribution of *p(x,y)*, over a set of random variables $V{=}x \cup y$, can be represented by a product of distributions that represent a smaller set of the full variable set [31].

$$(x, y) = \frac{1}{z} \prod_{a \epsilon F} \Phi_a(x_a, y_a) \tag{2.1}$$

Where, a is a subset of V

$$(F = a \subseteq V) \tag{2.2}$$

$$x =< x1, x2, , xn > \tag{2.3}$$

is the set of input variables for instance a sequence of tokens and

$$y =< y1, y2, .., yn > \tag{2.4}$$

is a set of output variables which for our case are the corresponding cause, effect or out tags for the tokens in a sentence. And Z defined in Eq. (2) is a constant that normalize Eq. (1) distribution to one.

$$Z = \sum_{x,y} \prod_{a \epsilon F} \Phi_a(x_a, y_a) \tag{2.5}$$

where

$$\Phi(x_a, y_a) = exp \left\{ \sum_{x,y} \lambda_{ak} f_{ak}(x_a, y_a) \right\} \tag{2.6}$$

The weights will be learned in a training procedure to positively reinforce the feature functions that are correlated with the output labels or assign negative values to feature functions that are not correlated with the output labels and zero values to uninformative feature functions.

For named entity extraction, MALLET [32] provides tools for sequence tagging. It makes use of algorithms like Hidden-Markov Models, Maximum Entropy Markov Model and Conditional Random Fields. To train the CRF model, data is manually annotated to form a training set. A validation set is used to verify the performance of the trained model. The model is trained till the time an increase in performance is noted. If there is a decrease in performance, the training will be stopped the model is tested on the test set to evaluate the model over unknown data. The CRF Model trains on the features of the text that is being analyzed. An example of a feature used to train the CRF model is a Parts-of-Speech tag of the text. The POS tag gives a lot of information about the structure of the text or sentence that is being analyzed.

Conditional Random Fields are a probabilistic framework for labeling and segmenting structured data. The work done by [33] presents a comparison study between CRFs and MEMMs and show that when both the models are parameterized in the exact same way, CRFs are more robust to inaccurate modeling than MEMMs. CRF also resolves the label bias problem which affects the performance of MEMMs. They also performed a POS tagging experiment where in CRFs performed better than MEMMs.

Several systems are using the CRF model for classification and prediction. [34] presents a system for the identification of sources of opinion, emotions and sentiments. They make use of CRF and a variation of AutoSlog [35]. This has been implemented in a two-fold fashion where-in the CRF module performs a sequence tagging and AutoSlog learns these extraction patterns. The CRF model is trained on three features which are three properties of the opinion source.

- The sources of opinions are mostly noun phrases.
- The source phrases should be semantic entities that can bear or express opinions.
- The source phrases should be directly related to an opinion expression.

The CRF model was developed using the code provided my MALLET. They also pointed out some errors due to sentence structure and limited vocabulary. The resulting system identified opinion sources with a precision of 80% and a recall of 60%.

Named entity recognition in Biomedical research is a most basic text extraction problem [36]. A Mallet based CRF model has been used for a machine learning system for NER. This method gives up to 85% precision and 79% recall for NER. [37] trains the CRF model using Orthographic features and Semantic features for named entity recognition. This framework was developed for simultaneously recognizing occurrences of PROTEIN, DNA, RNA, CELL-LINE, and CELL-TYPE entity classes. It was able to produce a precision and recall of 70%. Mallet based CRF has also been used to build a system that learns contextual and relational patterns to extract relations. In the work shown in [38], the CRF model was used for Parts-of-Speech tagging and was trained with sentence that contains relations and 53 labeled relations to extract relations from text. This method produced a precision and recall of 71% and 55% respectively. The use of CRF has also been done in discriminative part-based approach for the recognition of object classes from unsegmented cluttered scenes [39].

2.2 Summary

This chapter discussed the related work that has been put into the causal extraction, semantic tagging and conditional random fields. The techniques used for extraction varied from one approach to the other by the data source used and the method(s) involved in the process. It is evident that the structure of a sentence played a major role in the identification, classification or prediction of data. May it

be a deterministic or a probabilistic model, the problems arise from complex sentence structures. It can also be noticed from the examples citied that the approaches have been applied either in a single fashion or coupled with another method. The latter yielded better results as it had a higher level of refinement compared to the former. Implementing multiple information extraction processes into one system reduces the overall noise providing good quality results.

The next chapter presents the design and implementation of a multi-layered approach. Causal extraction techniques based on dictionaries has been used as bag-of-words. Semantic tagging has been implemented to enhance the use of the bag-of-words and conditional random fields have been implemented to identify actors or signals in the sentences.

3 DESIGN AND IMPLEMENTATION

3.1 Overview

The goal of this research is to develop a system that extracts causal sentences from the geriatric literature that have been fetched from Pubmed. When a causal sentence is detected, it is also important that the actors in the sentence are also detected.

All NLP systems work on a systematic approach. Figure 3.1 shows the process that we have applied for causal extraction. The causal mining approach starts by separating the Pubmed abstracts into sentences. Then tagging these sentences using Parts-of-Speech tagger and extracting a tag triplet that contains the semantic tag and marking the keyword in the triplet with the corresponding semantic tag. After the semantic tagging, the sentences with the right actors are to be identified. In order to understand the actors in a causal sentence, it is necessary that we analyze the different objects in a causal sentence and build a training model to identify similar actors in new sentences. For our purpose, the training model is built using conditional random field (CRF) which makes use of certain features of the words/phrases in the sentence. These features include the POS tag of the word and the shallow parser tags, which give us the information that the word is a part of a noun phrase or a verb phrase etc. Once the CRF model is trained, a new sentence is passed through the model for actor identification. Based on the actors identified, the sentence is classified into causal or non-causal.

3.2 Approaches for Causal Association Extraction

During the process of finding a solution to the causal extraction problem for geriatric literature, a number of conventional methods of classification and identification were used. These methods have been used by various other applications for natural language processing.

3.2.1 Naive Bayes Classifier Approach

Naive Bayes is a probabilistic classifier that is based on the Bayes Theorem [40]. We made of use of this method to classify causal and non-causal sentences from geriatric abstracts.

3.2.1.1. Method for Classification

The Naive Bayes classifier is trained for all sets for which classification is required. We trained the classifier with causal and non-causal sentences and tested the model on a fresh test set. We used a tool called Lingpipe [41] that provides a classification facility that takes samples of text classifications that are typically generated by an expert, and learns to classify new documents using what it learned with the language models.

The domain experts manually classified the sentences from the three categories, Fall Risk, Incontinence and Cognition into causal and non-causal sets. These sets were used to train the Naive Bayes classifier model and tests were performed in two strategies.

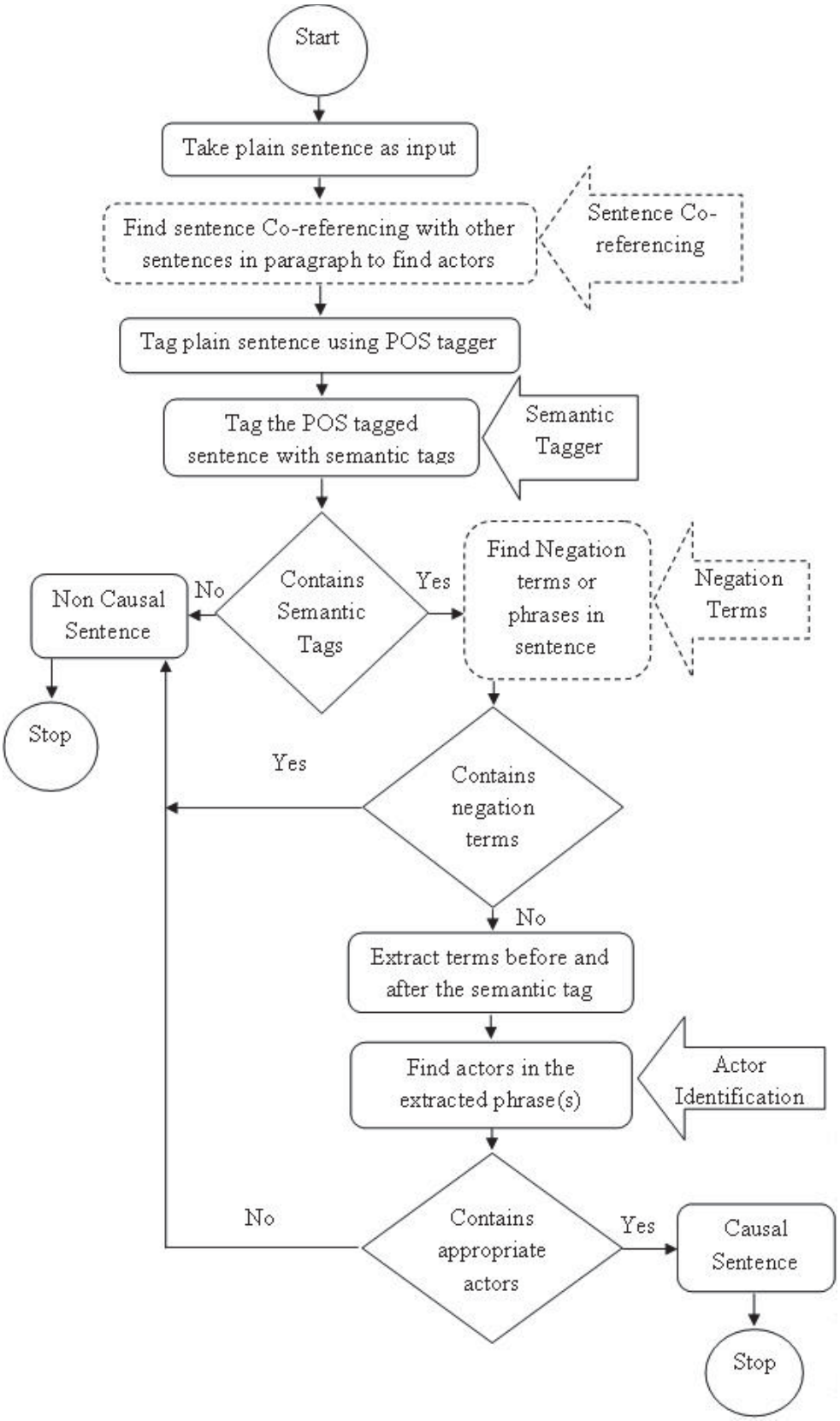

Figure 3.1.: Causal Extraction Process

3.2.1.1.1. Combinatorial

In combinatorial strategy, the aim was to determine which care-category has a higher coverage than the other sets, that is, which care-category is more comprehensive than other domains. In this approach, training set belonging to a single care-category is used on the test sets of all the domains. The results obtained were compared to see which domain gave the best accuracy. Table 3.1 shows the training and testing scenarios.

Table 3.1: Combinatorial strategy

Test Iteration	Training Set	Test Set
1	Fall Risk	Fall Risk
2	Fall Risk	Incontinence
3	Fall Risk	Cognition
4	Incontinence	Fall Risk
5	Incontinence	Incontinence
6	Incontinence	Cognition
7	Cognition	Fall Risk
8	Cognition	Incontinence
9	Cognition	Cognition

3.2.1.1.2. Cumulative

In cumulative strategy, a starting training set of a particular care-category (say Fall Risk) was used to run for each of the test sets in the three care-categories (Fall Risk, Incontinence and Cognition). After each test set is tested, the training set is retrained with the results from the previous test run. The results obtained were compared to see which training set would give the best accuracy. Table 3.2 shows the training and testing scenarios.

Table 3.2: Cumulative strategy

Starting Training Set	Test Set	Retrain	Test Set	Retrain	Test Set
Fall Risk	Fall Risk	->	Incontinence	->	Cognition
Fall Risk	Incontinence	->	Cognition	->	Fall Risk
Fall Risk	Cognition	->	Fall Risk	->	Incontinence
Incontinence	Fall Risk	->	Incontinence	->	Cognition
Incontinence	Incontinence	->	Cognition	->	Fall Risk
Incontinence	Cognition	->	Fall Risk	->	Incontinence
Cognition	Fall Risk	->	Incontinence	->	Cognition
Cognition	Incontinence	->	Cognition	->	Fall Risk
Cognition	Cognition	->	Fall Risk	->	Incontinence

It was noticed that at the end of the testing scenarios, the cumulative training set would be a summation of the training data from the three care-categories. There were some other factors that affected the performance of this approach. The factors are:

- Number of sentences in training set for each category,
- Length of sentences in the training set.

The results of this approach are explained in Chapter 4.

3.2.2 N-Gram based Approach

To overcome the problems that were identified in the Naive Bayes approach, we proposed a statistical approach to provide a simpler means to measure the probability using the N-Gram model. This method provides a probabilistic approach to analyze

and rate any term in the domain literature based on the number of occurrences of that term and to analyze the Parts-Of-Speech structure the term is present in. The text that is being analyzed has a considerable amount of common patterns that when extracted can be used for machine learning.

3.2.2.1. Method for Causal Extraction

After careful analysis of the sentences that were reviewed by the domain experts, it was found that each causal sentence comprises a phrase or term that makes that particular sentence, causal. For example,

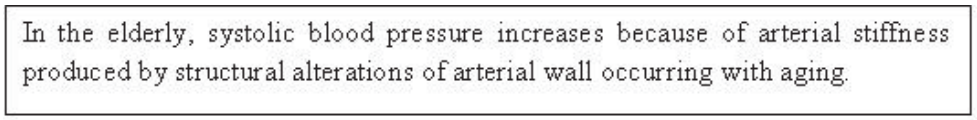

Figure 3.2.: Example of Causal Sentence

This is a causal sentence, Figure 3.2, which shows the relation between systolic blood pressure and arterial stiffness using the phrase increases because of. These relations are mainly defined by the existence of such key-phrases (or keyterms) and relation words. In some cases, the existence of relational words and the keywords does not mean that the sentence is causal. For example,

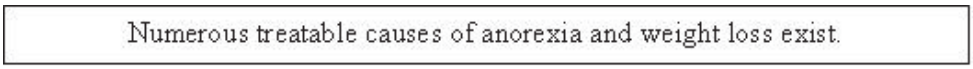

Figure 3.3.: Example of Non-Causal Sentence With Causal Term

In Figure 3.3, even though the term “causes” is present, the sentence still does not qualify as a causal sentence. The relational words do not always appear as a keywords or key-phrases. The sentences that do not contain such a relationship are termed Non-Causal. For example, in Figure 3.4,

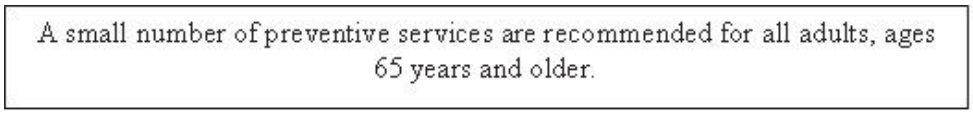

Figure 3.4.: Example of Non-Causal Sentence

This sentence does not exhibit the qualities of a causal relationship and is therefore classified as Non-causal.

Detection of the keywords is a Named Entity Recognition (NER) task. NER is a technique that finds the token boundary and the semantic category for particular terms occurring in the text. There are different approaches to NER. We used a dictionary approach to identify the keywords/key-phrases based on the review of a domain expert.

3.2.2.2. Building a Keyterm Dictionary

Once the terms or phrases are extracted, they are put into a table to form a keyterm dictionary. This can be explained with an example. Consider the following sentence, Figure 3.5, which has been marked Causal by the domain expert:

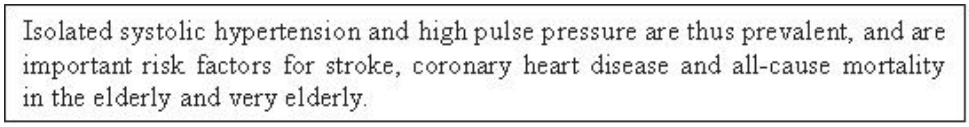

Figure 3.5.: Example of Non–Causal Sentence

The Figure 3.6 shows the structure of a causal phrase extracted from this sentence. The keyterm in this sentence is "risk factors". The value of N in the N-gram approach can be assigned only after analyzing various phrases from causal sentences.

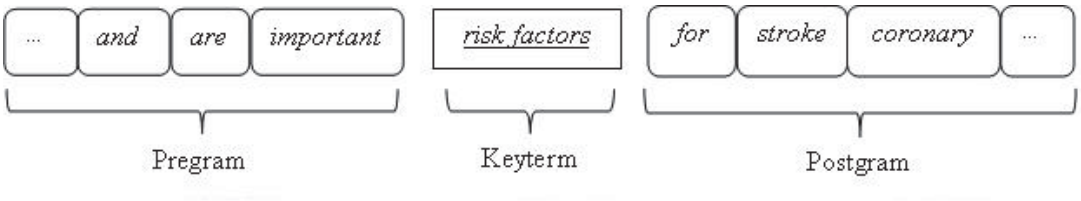

Figure 3.6.: Structure of Causal Phrase

3.2.2.3. Choosing the value of N for the N-Gram model

It was found that the word surrounding the central keyterm adds to the weight of the causal sentence.

We conducted several experiments after collecting keyterms on 1000 sentences to choose an appropriate value for N. The tests were run on one a randomly chosen category in the Geriatric domain. The results are shown in Table 3.3 and Figure 3.7.

Table 3.3: Specificity and Sensitivity to Choose Value of N

N-Value	True Positives	True Negatives	False Negatives	False Positives	Specificity (%)	Sensitivity (%)
N=1	19	136	11	13	91.27	63.33
N=2	23	128	7	21	85.9	76.67
N=3	24	120	6	29	80.53	80
N=4	25	119	5	30	79.86	83.33
N=5	25	115	5	34	77.18	83.33

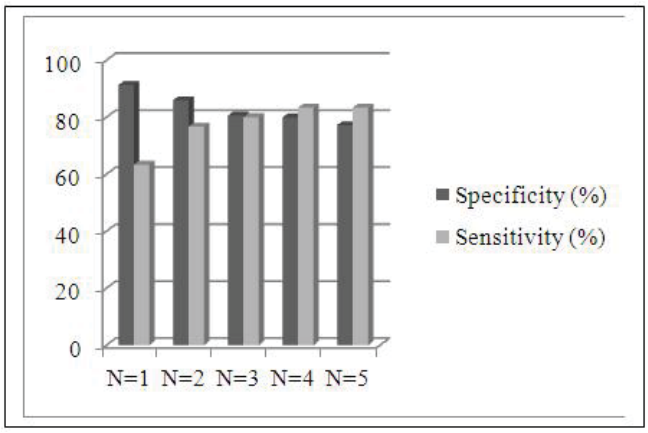

Figure 3.7.: Specificity and Sensitivity to Choose Value of N

We found that the optimal for N = 3, where N is the number of pregram and postgram terms, the system would provide optimum results.

Considering 3 pregram terms and 3 postgram terms to this keyterm, we have (Figure 3.8),

Pregrams terms: and, are, important

Postgrams terms: for, stroke, coronary

Figure 3.8.: Pregram and Postgram Terms

Analyzing over 19725 sentences, we have extracted 86 keyterms with 57 pregram and 23 postgram terms. Each of these terms is put together into separate dictionaries (along with their frequency of occurrence) called the keyterm dictionary, the pregram dictionary and a postgram dictionary. Table 3.4, Table 3.5 and Table 3.6 illustrate the various dictionaries.

The reason for creating three separate dictionaries is that the content of the pregram or the postgram terms is very different from the keyterm dictionary. The keyterms are specific to the domains whereas the pregrams and postgrams are words that are commonly used but influence the keyterms and thereby, the sentence.

Table 3.4: PRE-gram Word List

Term	Frequency	Score	Word Position
a	1	0.00869	PRE
additional	5	0.04347	PRE
adverse	1	0.00869	PRE
also	1	0.00869	PRE
are	3	0.02608	PRE
as	1	0.00869	PRE
be	1	0.00869	PRE
can	13	0.11304	PRE
claimed	1	0.00869	PRE
consequence	1	0.00869	PRE
could	1	0.00869	PRE
demonstrated	1	0.00869	PRE
depression	1	0.00869	PRE
disease	1	0.00869	PRE
dramatically	1	0.00869	PRE
effective	3	0.02608	PRE
greatest	1	0.00869	PRE
has	1	0.00869	PRE
have	5	0.04347	PRE
help	2	0.01739	PRE

3.2.2.4. Scoring the Terms

As we extract more and more keyterms, we also gather the frequency of occurrence of each keyterm in our sentence set. This gives us a clear idea of the significance of that keyterm as to how often do sentences with that word fall into the causal category.

Table 3.5: Keyword List

Term	Frequency	Score	Word Position
a promising intervention	2	0.0007434	KT
account	10	0.0037174	KT
affect	33	0.0122676	KT
aggravate	1	0.0003717	KT
alter	12	0.0044609	KT
an impact	2	0.0007434	KT
associated	350	0.1301115	KT
association	38	0.0141263	KT
attenuated	3	0.0011152	KT
attribute	5	0.0018587	KT
be associated	12	0.0044609	KT
be beneficial	2	0.0007434	KT
benefit	10	0.0037174	KT
carries	1	0.0003717	KT
cause	174	0.064684	KT
causing	12	0.0044609	KT
characterized	4	0.0014869	KT
complicate	1	0.0003717	KT
connected	1	0.0003717	KT
consequence	12	0.0044609	KT

For example, it was found that the term “associated” was present in about 313 causal sentences and the term “due to” was found about 56 times. Each of these

Table 3.6: POST-gram Word List

Term	Frequency	Score	Word Position
aging	1	0.02439	POS
between	3	0.07317	POS
a	1	0.02439	POS
depression	1	0.02439	POS
diminished	1	0.02439	POS
critical	1	0.02439	POS
role	1	0.02439	POS
by	3	0.07317	POS
for	2	0.04878	POS
have	1	0.02439	POS
in	1	0.02439	POS
increase	1	0.02439	POS
of	6	0.14634	POS
on	1	0.02439	POS
presence	1	0.02439	POS
significant	1	0.02439	POS
significantly	3	0.07317	POS
substantial	2	0.04878	POS
substantially	2	0.04878	POS
that	1	0.02439	POS

terms is assigned scores based on its frequency in each individual dictionary. The score of the keyterm can be given using an expression as

$$X_K T = \frac{\sum_{i=0} n\omega_i}{\omega_i} \tag{3.1}$$

Where

X is the score assigned.

ω is the frequency of the keyterm (KT) in a dictionary.

And n is the number of keyterms identified from the sentence set.

The scores are assigned to terms of all the 3 dictionaries. Once the scores are assigned, the dictionaries are used as a basis for identifying causal sentences from any fresh abstract or article.

When a sentence was processed, and a keyterm is identified, the pre and the post grams of that keyterm from the sentence are extracted and matched with the dictionaries. The corresponding scores of the terms are multiplied to come up with a score for that sentence. All the sentences in the test set were put through this process and a threshold was set to filter out the causal sentences. The results and problems that came out of this approach are explained in Chapter 4.

During the testing of the N-gram approach, there were some problems that were detected.

- The frequency of occurrence of some of the strong and weak causal terms had been found to be similar or very close to each other which did not play a fair role in identifying strong causal sentences.
- The keyterm dictionary, although was built using the care-categories, did not converge and was unable find causal sentences from care-categories.

Due to these reasons, the performance of the system was not satisfactory. The results of this approach are explained in Chapter 4.

3.3 Methodology for Multi-layered approach

3.3.1 Semantic Tag Extraction from Literature

Semantic tagging, as explained in section 2.2.3, is a method of assigning tags or markers to text strings which can help in making them discoverable. Implementing a Semantic Tag for any domain needs careful understanding of the domain. This is because the sentence structures and the word usage differ from domain to domain.

This section discusses a kind of semantic tagging that tags a sentence based on the Parts-of-speech of a phrase and the terms that the phrase contains.

3.3.1.1. POS Tag triplets

Causal sentences in documents have varying forms as found in any natural language based text. There are several ways of detecting these causal sentences.

- The straight forward approach is to find the occurrence of the keyword in the sentence using simple string matching algorithm. In this case, all possible forms of the keyword can be extracted.
- Another method is to apply a syntactic tag to the sentence and find a syntactic tag sequence and then check for the occurrence(s) of the keyword. This will restrict the detection of the keyword only if the keyword occurs in a certain form.

Detecting a particular form of the keyword, based on the POS tag sequence of the keyword helps in fine tuning the causal keyword search to reduce the noise that occur using straight forward approaches.

In the example shown in Figure 3.9 and Figure 3.10, we can see that a causal keyword can be used in multiple forms and the same causal term occurs in causal as well as non-causal sentence but the POS tags that they are coupled with are different.

However, further research should focus on the optimum dose *relationship* of frequency, amplitude and duration for the various populations.

Figure 3.9.: Causal Term in Non-Causal Sentence

Mobility disability is one of the major *risk factors* for morbidity and mortality in this age group.

Figure 3.10.: Causal Term in Causal Sentence

In this approach, the text is preprocessed by providing syntactic tags in the form of POS tags. Then the POS tags are extracted from the text in groups of three to form the POS Tag triplets. All these tag triplets are collated and stored in a table.

The POS tag triplets are extracted using the following approach shown in Figure 3.11 and Figure 3.12.

1. Apply POS tagging to all the sentences.
2. Search for the causal keyterms in the POS tagged sentences.
3. Once a match is found, extract the causal term, one term before the causal term and one term after the causal term along with the POS tags from the tagged sentence.
4. Extract the three POS tags from the phrase in step 3.
5. This forms the POS tag triplet.

Figure 3.11.: POS Tag Triplet Extraction Approach

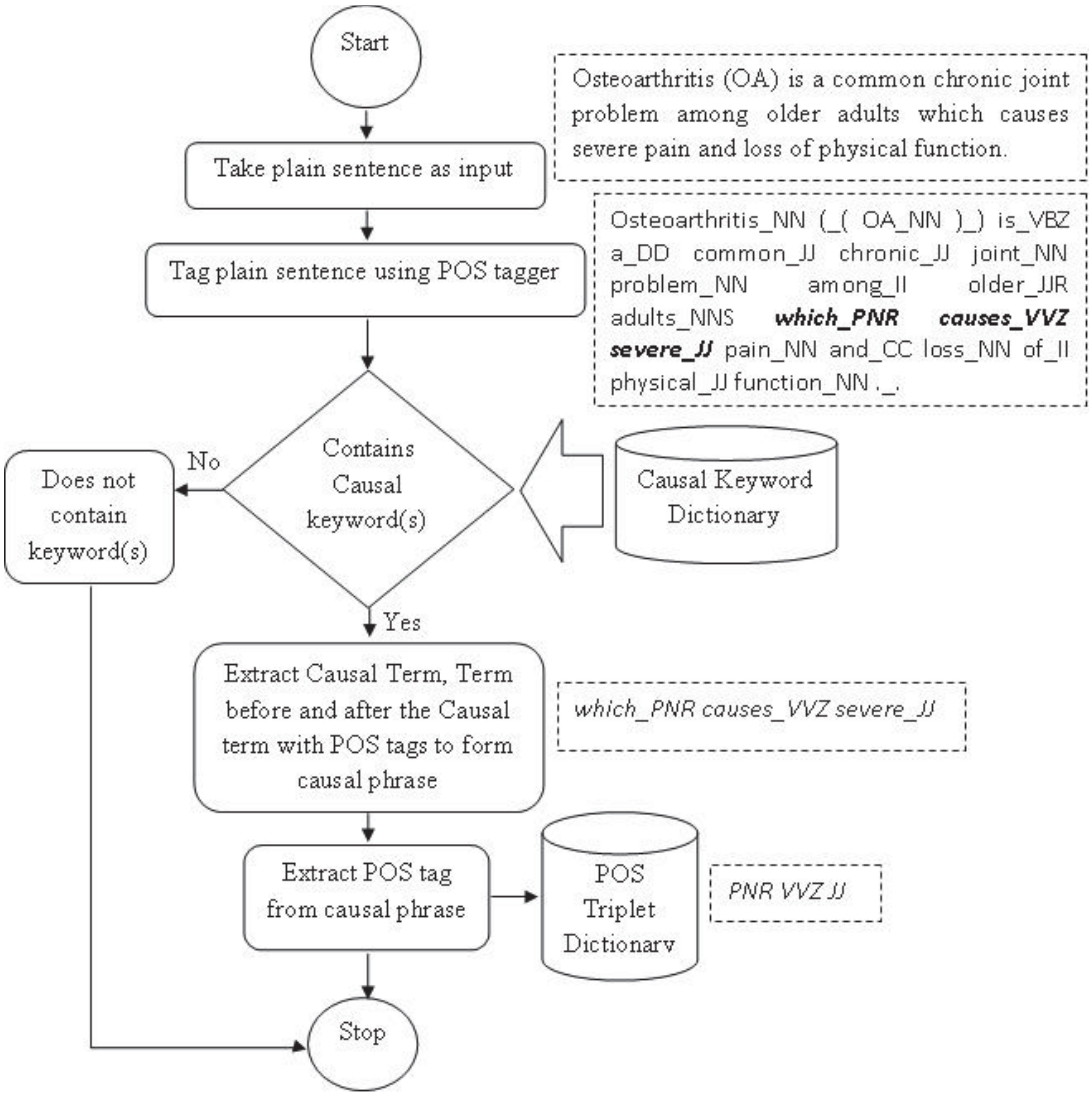

Figure 3.12.: POS Tag Triplet Extraction Process

After extracting the POS tag triplets from the sentences and careful analysis led to an observation that if the individual components of the triplets were collated, there exists a mapping of the POS tags which is shown in Figure 3.13.

This mapping can be used on other domains to see the patterns of the occurrence of POS tags along with the keyterms in that domain.

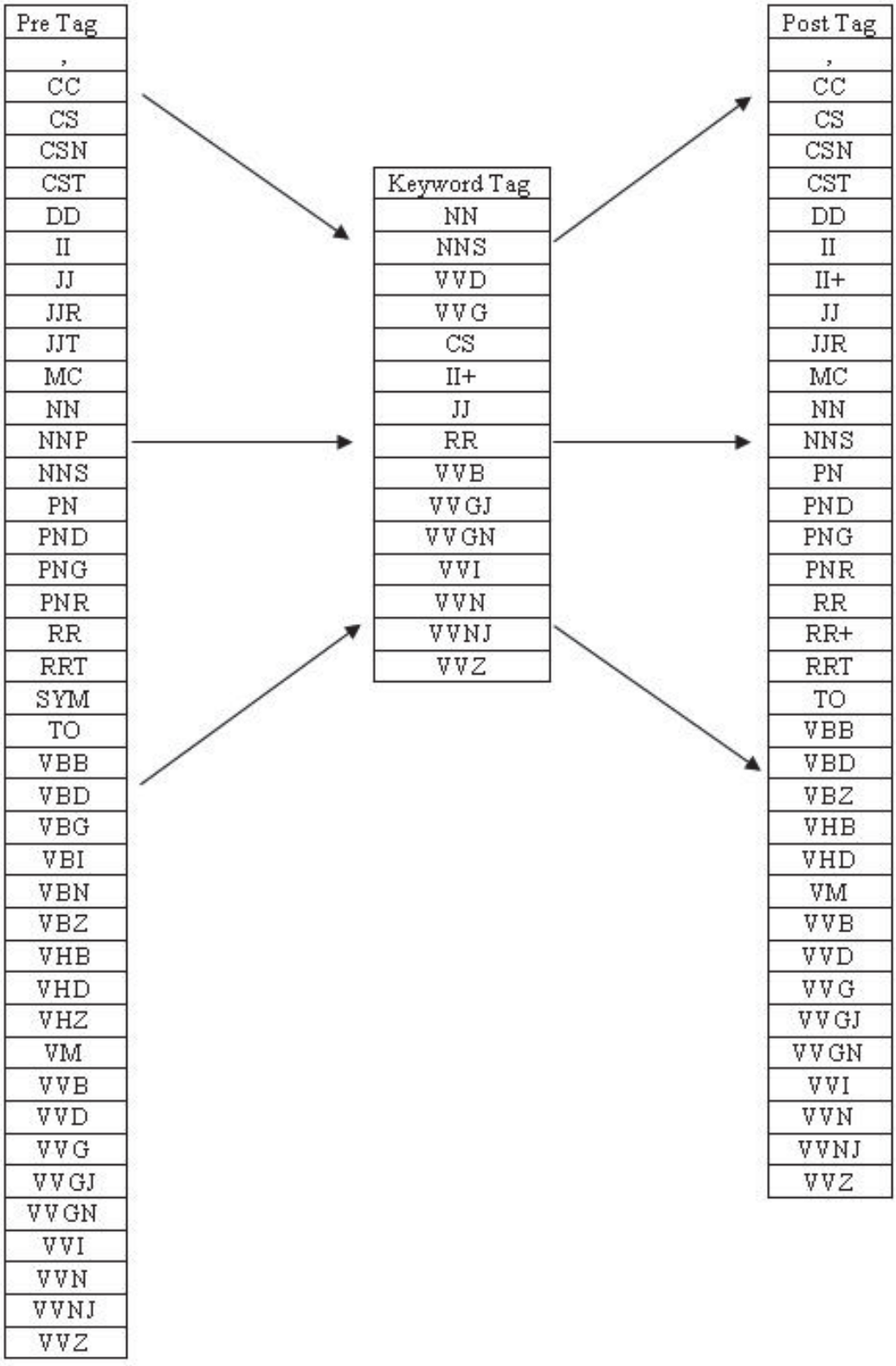

Figure 3.13.: POS Tag Triplet Mapping

3.3.1.2. Causal Keyterms

The keyterm table shown in Table 3.5 contains most of the causal keywords that have been extracted from the geriatric abstracts. This table was used as the start up dictionary for constructing the semantic tag table. As more care categories were added as part of the research, more causal keyterms were discovered to form a new keyterm dictionary. The new dictionary has a more comprehensive set of causal keywords that is used for causal extraction.

3.3.1.2.1. Semantic Groups

In the N-gram model, the keyterm dictionary was used to find an occurrence of the causal keyword in the sentences. The extracted sentences that were analyzed provided no information about the relationship that was extracted from the sentence. For example,

Sentence 1: Osteoarthritis (OA) is a common chronic joint problem among older adults which *causes* severe pain and loss of physical

Figure 3.14.: Causal Sentence With "cause" Keyword

Sentence 2: Objectively measured light-intensity physical activity is *associated* with physical health and well-being variables in older adults.

Figure 3.15.: Causal Sentence With "associated" Keyword

Sentence 3: Alzheimers disease *results* in the cognitive and functional deterioration of the affected patient, and behavioral disturbances frequently accompany the disease.

Figure 3.16.: Causal Sentence With "result" Keyword

Sentence 1 in Figure 3.14, contains a “cause” relation, sentence 2 in Figure 3.15, contains an “association” relation and sentence 3 in Figure 3.16, contains a “result” relation between the entities in the sentence.

To get a better understanding of the nature of the sentences that are extracted as causal, the causal keywords have to be arranged into groups with a name or tag assigned to each group. This approach was applied to the keyword dictionary and the causal keyterms were divided into 9 different groups. Table 3.7 shows the 9 groups and the tags assigned to each of these groups.

The idea behind semantic tagging is to make information discovery as efficient and refined as possible. This means that the extracted information should be meaningful and logically correct. The approach in this research describes the use of the POS tag triplets and the semantic groups.

3.3.2 Extracting Keyphrase from Text

The semantic tag generated in the approach above does not uniquely match a single keyphrase. The usage of many causal keywords has similar POS tag triplet patterns. For example Figure 3.17 shows a causal phrase where the POS tag triplet extracted from the phrase is DD NN II which is same as the pattern extracted from Figure 3.18.

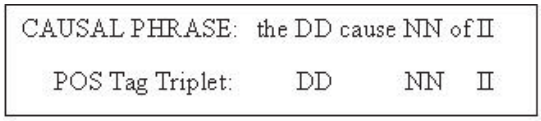

Figure 3.17.: Causal Phrase With “cause” Keyword and POS Triplet

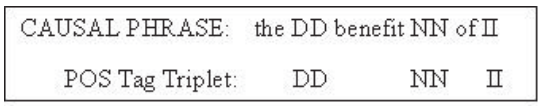

Figure 3.18.: Causal Phrase With “benefit” Keyword and POS Triplet

Table 3.7: Semantic Groups

TASO	TCAU	TCON	TDEC	TEFF	TINC	TIND	TRES	TOTH
associated	cause	contribute	reduce	effect	exacerbate	indicative	result	carries
association	causing	contributing	inhibited	affect	exasperate	indicator	resulting	experience
correlated	due	contribution	decrease	impact	enhance	identify	results	incidence
correlation	because	contributors	degrade	predict	enhancing	tended	resulted	problem
linked	create	facilitate	deteriorates	influence	evokes	characterized	consequence	resolving
interaction	factor	account	hinder	effects	increase	likelihood	costs	risk
relate	induce	plays	impair	affected	aggravate	likely	noted	occur
relationship	caused	attributed	mitigate	influenced	encourage		suggest	complicate
connected	causes	contributes	reducing	hampered	promote		leads	allowed
mediate	induced	contributed	reduction	affects	benefit		leading	
constitutes	galvanize	attributable	minimize	influencing	help		resulting	
associations	originate	imposes	attenuated	impacts	helpful		lead	
mediated	derail		decline	influences	perpetuate		resulted	
connecting	arouses		falling	predictors	attribute		implicate	
related	factors		precipitate	predicted	improve		consequences	
	impinge		prevent	predictor	improving			
	pose		reduces	disrupt	enhanced			
			reduced	disruptive	increases			
			prevents	disruption	improves			
			eliminate	disruption	increased			
			precipitated					
			decreasing					
			decreases					

Even though the keyterms involved in the casual phrase are different, they provide similar POS tag patterns. This gives the system the ability to have a finite set of tag triplets that will be searched for in the sentence.

The process of semantic tagging of the text can be explained using the method in Figure 3.19 and illustrated in the Figure 3.20.

a. First a sentence is fetched from the sentence set.
b. POS tagging of the applied on the sentence.
c. Extract only the POS tag sequence from the sentence.
d. Search the POS tag sequence for a POS tag triplet.
e. If a POS tag sequence is found, extract the corresponding text phrase from the sentence.
f. Search the text phrase and check if it contains a causal keyword.
g. If a causal keyword is found, replace the causal keyword with the semantic tag.
h. Store the sentence with the semantic tag back into the sentence set.

Figure 3.19.: Approach for Semantic Tagging

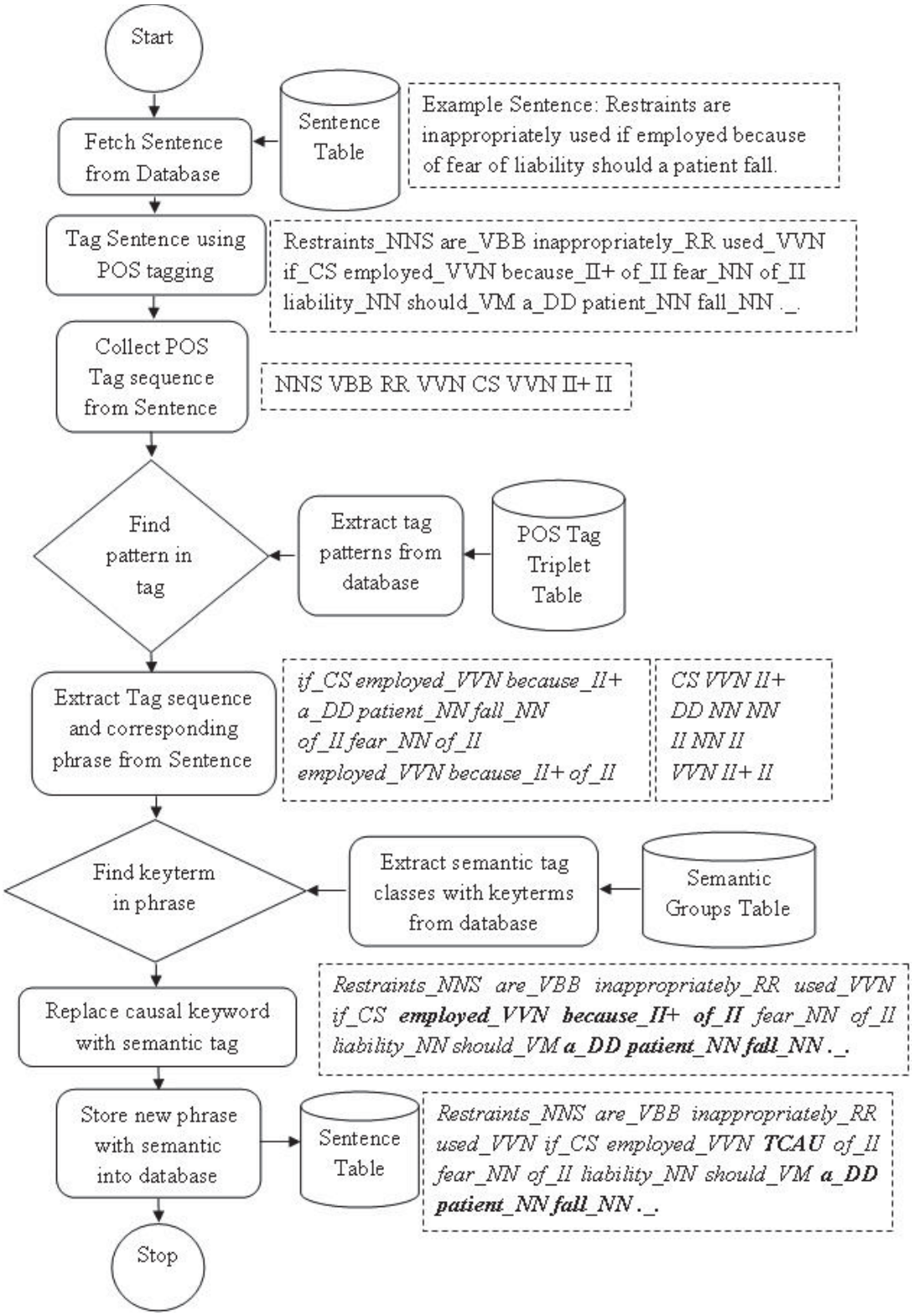

Figure 3.20.: Semantic Tagging Approach

3.3.3 Creation of Semantic Tags for Geriatric Domain

Extraction of causal sentences from geriatric literature involves the use of a combination of the POS tag triplets and the semantic groups to form a semantic tag.

Figure 3.21 shows the table that stores all the combinations of PRE (pre keyterrn), KT (Keyterm) and POS (post keyterrn) in the geriatric text. The semantic groups shown in Table 3.18 are also stored in a database table for use in causal extraction.

Once the POS tag triplets and the semantic groups are formed, they are together used to form the semantic tag. Figure 3.21 shows an illustration of how a semantic tag is formed.

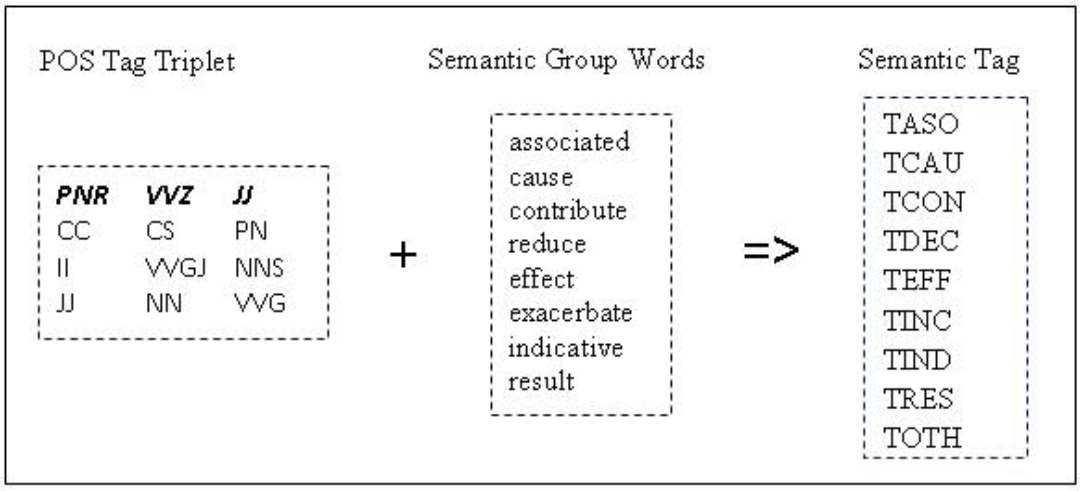

Figure 3.21.: Formation of Semantic Tag

Considering the example shown in Figure 3.20, the phrase *"employed_VVN because_II+ of_II"* is converted to *"employed_VVN TCAU of_II"* suggesting the TCAU tag being identified in the sentence.

3.4 Actors in Geriatric Literature

Actor identification is another challenging task in a natural language processing system. This involves domain experts to analyze documents, identify the important

entities in the sentence and annotate the text with these entities. The idea behind actor identification is the concept of signal identification discussed in section 2.1.3.

The geriatric care domain has several actors that are used specifically in the geriatric care literature. Since some of the actors can be found in other health care related documents, developing a method for identifying these actors in the geriatric care can be used in other related domains as well.

We use a learning model to find actors in geriatric sentences that is coupled with the causal extraction process.

3.4.1 Identifying Actors in Sentences

Actor is a term used to indicate presence of specific information. In causal sentence, the presence of an actor(s) shows the relation between the entities and also the kind of relation it is involved in. A lot of dependence is placed on the context of use of the actor. Depending on the context, the actor may or may not be obvious.

In this research work, Conditional Random Field (CRF) has been used as the learning model. The important reason for the use of CRF over other learning models is due to the following reasons.

- Support Vector Machine requires a large amount of training and testing data.
- CRFs avoid the label bias problem, which is a weakness shown by Maximum Entropy Markov Models (MEMMs) and other conditional Markov models.

3.4.2 Conditional Random Fields

Conditional Random Fields, as discussed in section 2.1.4, is a probabilistic framework for labeling and segmenting structured data. This section explains how CRF is used and implemented for actor identification.

3.4.2.1. CRF Features

Features are inputs to the CRF model and the outputs are a sequence of actors and out tags. In the training and validation set, we supply the CRF model with feature inputs and known outputs so that the CRF model can learn the pattern of data by adjusting the weights of its feature function. The learned model then will be used on an unseen data (test set) to predict its sequence of tags.

3.4.2.2. Creating Training Data

As in the case of any machine learning system, the CRF model has to be trained properly with a data set which has good instances of entities. The entities in our study relates to actors in a sentence. CRF requires either positive or negative instances of sentences. Since the leaning model is being constructed to identify actors, sentences with geriatric actors are annotated. We use the CRF method provided by Mallet. Mallet is package built on the Java platform for statistical natural language processing, classification, clustering, topic modeling, information extraction, and other machine learning applications. The Mallet based CRF model accepts training data in the format shown in Figure 3.22.

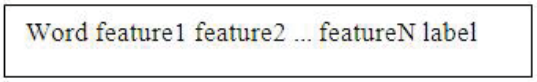

Figure 3.22.: Mallet Training Input Format

The actor(s) marked in a sentence are the labels given to that sentence. For the purpose of our research, the features chosen for the sentence are

- The POS tag of the words.
- The Shallow parser tags for the phrases.

A shallow parser performs chunking of words used in a sentence and identifies the constituents like the noun groups, verbs, verb groups etc. Shallow parser tags can be

very useful feature of a word in a sentence and can be used in combination with the POS tag of that word to enhance the training of the CRF model.

Once constructed, the training set would look like the example below. Figure 3.23 shows a sentence that needs to be annotated for actors. Table 3.8 shows the annotated sentence marked with words/phrases that are considered to be actors of the geriatric care domain.

Chronic obstructive pulmonary disease (COPD) is a debilitating disease of the elderly that causes significant morbidity and mortality.

Figure 3.23.: Sentence to be converted to Mallet Training Input Format

The CRF model is learned on the training set constructed by annotating causal sentence in the format shown in Table 3.8. The amount of training data is purely based on how clean the training set is. If the training set contains a lot of noise or redundant data, the training model thus created will be ambiguous leading to poor prediction and performance. On the other hand, if the training set is clean and accurate, the model will provide better performance.

3.5 Summary

This chapter described three major methods in the quest for extracting causal relations from geriatric sentence. The methods ranged from a probabilistic Nave Bayes method, to a deterministic N-gram model to a combination of Syntactic and Semantic tagging along with CRF model. Although any of the extraction models could have been used for our purpose, the performance of each method proved otherwise. The next chapter will discuss the experiments that were run on all these methods and the related results and performance that proved the hypothesis.

Table 3.8: Sample CRF Training Data

Sentence	POS Tag	Shallow Parser Tags	Actor/Non-Actor Labels
Chronic	JJ	B-NP	I-Actor
Obstructive	JJ	I-NP	I-Actor
Pulmonary	JJ	I-NP	I-Actor
Disease	NN	I-NP	I-Actor
(	(	O	O
COPD	NN	B-NP	I-Actor
)	)	O	O
Is	VBZ	B-VP	O
A	DD	B-NP	O
Debilitating	NN	I-NP	O
Disease	NN	I-NP	O
Of	II	O	O
The	DD	B-NP	O
Elderly	NN	I-NP	I-Actor
That	PNR	B-VP	O
Causes	VVZ	B-NP	O
Significant	JJ	I-NP	O
Morbidity	NN	I-NP	I-Actor
And	CC	I-NP	O
Mortality	NN	I-NP	I-Actor
.	.	O	O

4 EXPERIMENTS AND RESULTS

In the process of implementing the system, several experiments were conducted at every stage. The experiments were run on 42 care categories; total of 19725 sentences were put through this experiment to determine the performance of causal extraction after the implementation of individual modules to the research work. Section 4.2 of this chapter presents the performance of methods used for causal extraction. Section 4.3 shows the results after applying the semantic tags to the sentences. Section 4.4 contains the experiments performed on actor identification which the results of the learning model and testing the sentences on this model. In addition, section 4.5 shows the results of the testing the system on all the geriatric domains. Finally, section 4.6 shows the comparison of all the results shown on the research work so far.

4.1 Calculation of results

For evaluating the results we did precision, recall, false positive rate, f-score, and accuracy calculations [42]. The formulas used to calculate these values are given by Equation 4.1, Equation 4.2, Equation 4.3, Equation 4.4 and Equation 4.5.

$$Precision = \frac{TP}{TP + FP} \tag{4.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.2}$$

$$FalsePositiveRate = \frac{FP}{FP + FN} \tag{4.3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.4}$$

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4.5}$$

4.2 Performance of Causal Association Extraction Methods

4.2.1 Naive Bayes Performance

The results for the combinatorial approach are shown in Tables 4.1, Table 4.2 and Table 4.3.

Table 4.1: Performance - Fall Risk on Other Care-Categories

Training Set	Test Set	Total Sentences In Test Set	True Positives	True Negatives	False Negatives	False Positives
Fall Risk	Fall Risk	203	27	99	8	69
Fall Risk	Cognition	195	32	122	7	34
Fall Risk	Incontinence	190	23	99	13	55
	Precision	Recall	False Positive Rate	Accuracy		
	28.13%	77.14%	41.07%	62.07%		
	48.48%	82.05%	21.79%	78.97%		
	29.49%	63.89%	35.71%	64.21%		
Average	35.37%	74.36%	32.86%	68.00%		

The Fall Risk Training set consisted of 35 Causal sentences and 168 Non-Causal sentences. The experiments of the Fall Risk training set on the three sets were not able to provide the expected results. It was able to provide a Precision of 35.37%, Recall of 74.36% with a False Positive Rate of 32.86%.

Table 4.2: Performance - Cognition on Other Care-Categories

Training Set	Test Set	Total Sentences In Test Set	True Positives	True Negatives	False Negatives	False Positives
Cognition	Fall Risk	203	17	149	18	19
Cognition	Cognition	195	11	136	28	20
Cognition	Incontinence	190	10	138	26	16
	Precision	Recall	False Positive Rate	Accuracy		
	47.22%	48.57%	11.31%	81.77%		
	35.48%	28.21%	12.82%	75.38%		
	38.46%	27.78%	10.39%	77.89%		
Average	40.39%	34.85%	11.51%	78.00%		

The Cognition Training set consisted of 39 Causal sentences and 156 Non-Causal sentences. The experiments of the Cognition training set on the three sets were able to yield a Precision of 40.39%, Recall of 34.85% and a False Positive Rate of 11.51%. This training set also did not prove to be good enough to classify the sentences.

The Incontinence Training set consisted of 36 Causal sentences and 154 Non-Causal sentences. When run on the three sets, It was able to provide a Precision of only 36.41%, Recall of 59.09% and a False Positive Rate of 24.66%. The results for the cumulative approach given in Table 4.4.

Table 4.3: Performance - Incontinence on Other Care-Categories

Training Set	Test Set	Total Sentences In Test Set	True Positives	True Negatives	False Negatives	False Positives
Incontinence	Fall Risk	203	18	129	17	28
Incontinence	Cognition	195	22	114	17	42
Incontinence	Incontinence	190	25	109	11	45
	Precision	Recall	False Positive Rate	Accuracy		
	39.13%	51.43%	17.83%	76.56%		
	34.38%	56.41%	26.92%	69.74%		
	35.71%	69.44%	29.22%	70.52%		
Average	36.41%	59.09%	24.66%	72%		

The results for the Cumulative approach were done by collating the training set of all the training sets and testing them on individual sets. The "Whole Set" given in Table 4.4 contained 110 Causal sentences and 478 Non-Causal sentences. This approach also did not yield any satisfactory results with a Precision of 36%, Recall of 59.32% and a False Positive Rate of 24.74%.

From the results in the above tables, it can be seen that the Naive Bayes approach did not converge and hence did not provide good performance. Both the strategy to find an optimum training model did not yield good results.

Table 4.4: Performance - Whole Set on Other Care-Categories

Training Set	Test Set	Total Sentences In Test Set	True Positives	True Negatives	False Negatives	False Positives
Whole Set	Fall Risk	202	24	119	11	49
Whole Set	Cognition	195	21	128	18	28
Whole Set	Incontinence	190	20	113	16	42
	Precision	Recall	False Positive Rate	Accuracy		
	32.88%	68.57%	29.17%	70.44%		
	42.86%	53.85%	17.95%	76.41%		
	32.26%	55.56%	27.10%	69.63%		
Average	36.00%	59.32%	24.74%	72%		

4.2.2 N-Gram Performance

The keyterm dictionary was constructed using a subset of the care-categories and it was expected that other care-categories would have similar word usage and structure to fit into the keyterm dictionary. Figure 4.1 shows the performance of the system. The Precision was calculated to be 66%, Recall was 74% and the False Positive Rate was at 16%.

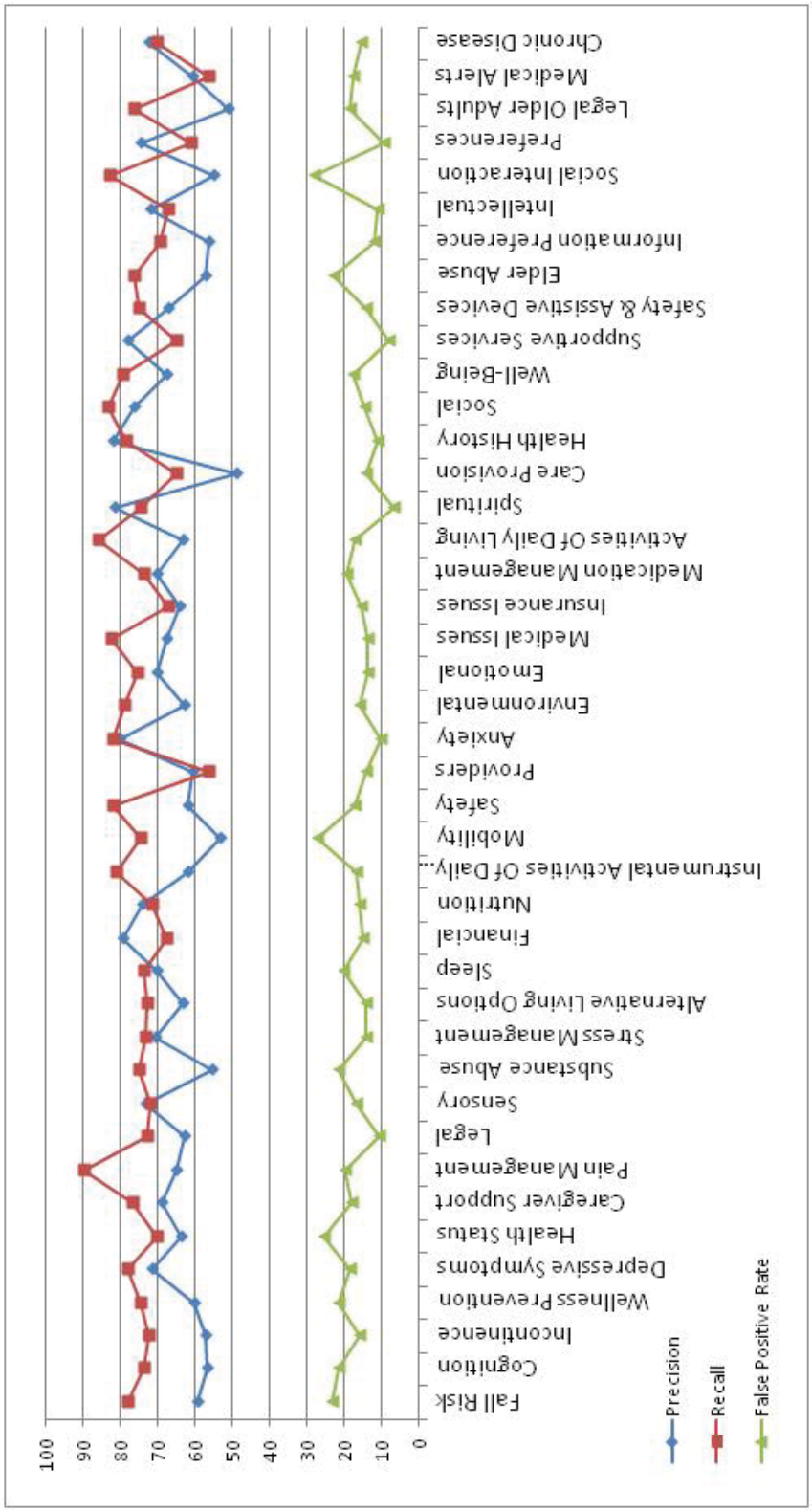

Figure 4.1.: Performance of N-Gram Approach

4.3 Semantic Tag Extraction

Semantic tag extraction was done in two steps.

4.3.1 Extraction of keywords from geriatric text

As explained in section 3.2.2.1.1 which talks about building a keyword dictionary, extracting keywords from the geriatric text is a manual process of constructing the keyword dictionary.

4.3.2 Extraction of POS Tag triplets

Extraction of the POS tag triplets is done using the approach given in section 3.3.1.1. Table 4.1, Table 4.2 and Table 4.3 illustrates an example of the step-by-step results of extracting POS tag triplets.

In the first step, the keyterm word is extracted from the POS tagged text with one word before and one word after the keyterm along with all the POS tags of the individual terms. Next, the keyterm, the pre-keyterm and the post keyterm words are removed from the phrase to obtain the three POS tags. The list extracted in step 2 can contain duplicates which are removed in the final step. This process is applied on all the keyterms and once the entire list is obtained, they are stored in tables to be used during semantic tagging.

4.4 Experiments on Applying Semantic Tags

Once the sentences were semantically tagged, as per the first part of the causal extraction process in Figure 3.1, if a sentence contains a semantic tag, it is marked, causal; if not then it is marked non-causal. For validation, a new set of sentence was identified that contained unknown abstracts from the geriatric domain. This set

Table 4.5: First Step of POS Tag Triplet Extraction – Extracting Pre word, Keyterm, Post Work and Tags

Pre Word	POS Tag	Keyterm Word	POS Tag	Post Word	POS Tag
Or	CC	Eliminate	JJ	Symptoms	NNS
uniformly	RR	Eliminate	JJ	USOC	NN
Even	RR	Eliminate	VVB	The	DD
To	TO	Eliminate	VVI	Or	CC
To	TO	Eliminate	VVI	inter-rater	JJ
To	TO	Eliminate	VVI	Oral	JJ
Should	VM	Eliminate	VVB	Acute	JJ
Can	VM	Eliminate	VVB	Tobacco	NN
reforms	VVZ	Eliminate	NN	Or	CC

contained 165 sentences and was manually classified by the domains experts. The tests were performed on the validation set and results are given in Table 4.4.

4.5 Experiments on Actor Identification

Actor identification was performed using CRF with the Mallet tool. This process involved creating a training data and generating a trained model to test the test set.

4.5.1 Training

CRF creates a model based on the features that are provided for learning. For this, a data set of 800 different instances of the sentences from the various geriatric

Table 4.6: Second Step of POS Tag Triplet Extraction – Removing words

Pre Word POS Tag	Keyterm Word POS Tag	Post Word POS Tag
CC	JJ	NNS
RR	JJ	NN
RR	VVB	DD
TO	VVI	CC
TO	VVI	JJ
TO	VVI	JJ
VM	VVB	JJ
VM	VVB	NN
VVZ	NN	CC

care-categories were chosen, annotated and used. Table 3.8 shows an example of the training set used by the CRF model.

4.5.2 Testing

Actor identification was performed only on those sentences that were marked as causal at the end of the semantic tagging procedure. The reason for doing this is that the aim of the research work is to identify causal sentences and only those sentences that contain a semantic tag can indicate causal behavior and hence can be used to identify actors. Once actor(s) are identified in a sentence, as per the final step of the causal extraction process in Figure 3.1, only those sentences that contain actor(s), are marked as causal; if not then they are marked non-causal.

The tests performed on the validation set proved to be an improvement to the results found by the semantic tagging. The results are given in Table 4.8.

Table 4.7: Third Step of POS Tag Triplet Extraction – Removing Duplicate Tag Triplets

Pre Word POS Tag	Keyterm Word POS Tag	Post Word POS Tag
CC	JJ	NNS
RR	JJ	NN
RR	VVB	DD
TO	VVI	CC
TO	VVI	JJ
VM	VVB	JJ
VM	VVB	NN
VVZ	NN	CC

Table 4.8: Performance of Semantic Tagging on Validation Set

Domain Name	True Positives	True Negatives	False Negatives	False Positives
Test Domain	42	99	10	13
Precision	Recall	False Positive Rate	Accuracy	F-Measure
76.36%	80.77%	11.60%	85.98%	78.50%

Table 4.9: Performance of Semantic Tagging and Actor Identification on Validation Set

Domain Name	True Positives	True Negatives	False Negatives	False Positives
Test Domain	37	99	9	9
Precision	Recall	False Positive Rate	Accuracy	F-Measure
80.43%	80.43%	8.33%	88.31%	80.43%

4.6 Testing and Validation with Sentences from All Geriatric Domains

Once the tests were performed on the validation set, it was partly confirmed that the system was capable of extracting causal sentences. The confirmation of the tests can be achieved only after executing the system on the all the care-categories and comparing the results across them. The results after executing the system on all the 42 care-categories are given in Table 4.10 and Figure 4.2.

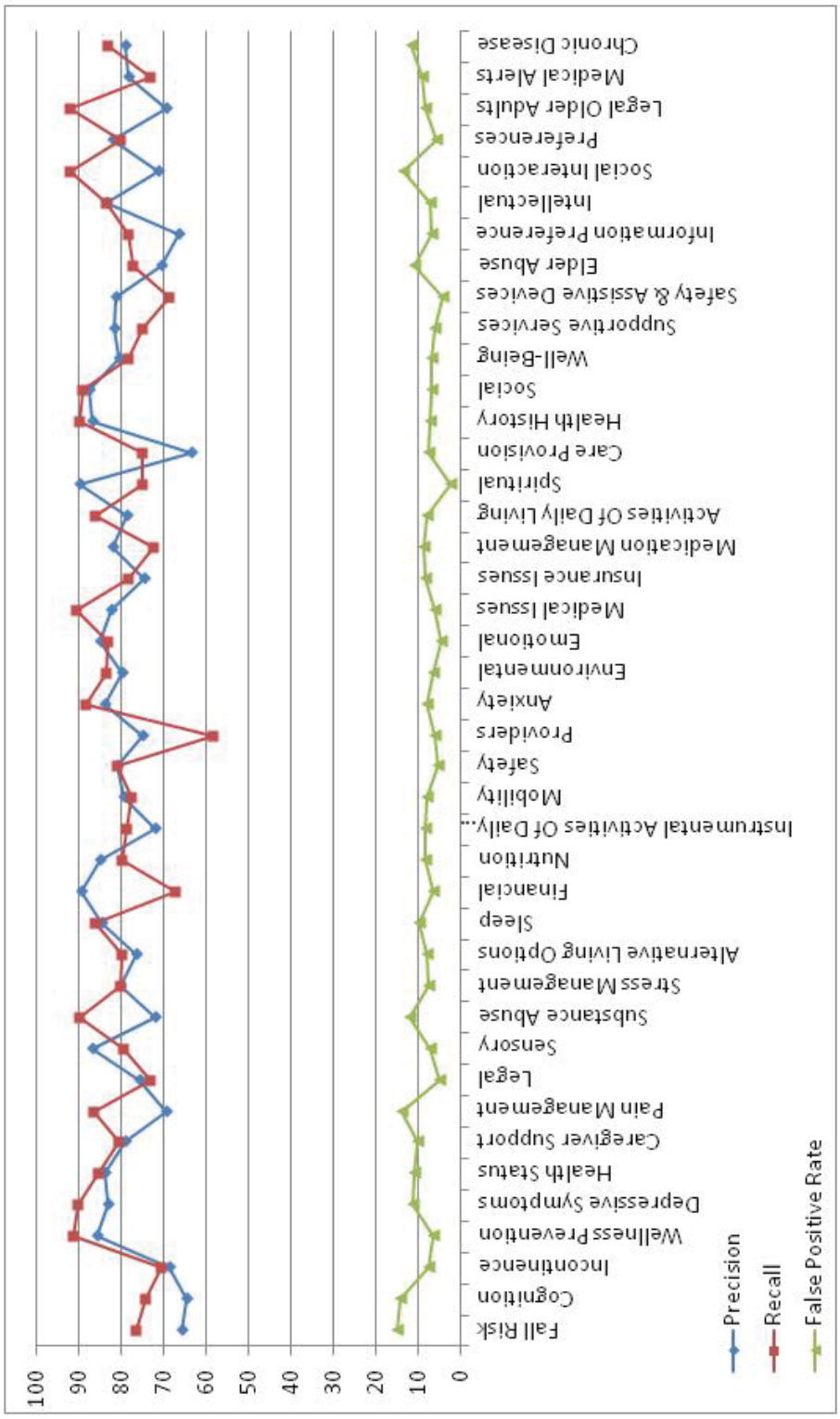

Figure 4.2.: Performance of Semantic Tagging and Actor Identification

Table 4.10: Performance of Semantic Tagging and Actor Identification on All Domains

Domain Id	Domain Name	True Positive	True Negative	False Negative	False Positive	Precision	Recall	False Positive Rate	Accuracy	F-measure
600	Fall Risk	782	2295	241	409	65.66	76.45	15.13	82.56	70.65
601	Cognition	623	2034	217	341	64.63	74.17	14.36	82.65	69.07
602	Incontinence	465	2488	194	211	68.79	70.57	7.82	87.94	69.67
603	Wellness Prevention	43	100	4	7	86	91.49	6.55	92.86	88.66
604	Depressive Symptoms	55	86	6	11	83.34	90.17	11.35	89.25	86.62
605	Health Status	58	88	10	11	84.06	85.3	11.12	87.43	84.68
606	Caregiver Support	42	95	10	11	79.25	80.77	10.38	86.71	80
607	Pain Management	32	86	5	14	69.57	86.49	14	86.14	77.11
608	Legal	22	133	8	7	75.87	73.34	5	91.18	74.58
609	Sensory	66	126	17	10	86.85	79.52	7.36	87.68	83.02
610	Substance Abuse	44	123	5	17	72.14	89.8	12.15	88.36	80
611	Stress Management	37	107	9	9	80.44	80.44	7.76	88.89	80.44

612	Alternative Living Options	36	127	9	11	76.6	80	7.98	89.08	78.27
613	Sleep	50	82	8	9	84.75	86.21	9.9	88.6	85.48
614	Financial	43	71	21	5	89.59	67.19	6.58	81.43	76.79
615	Nutrition	40	75	10	7	85.11	80	8.54	87.13	82.48
616	Instrumental Activities Of Daily Living(Iadls)	26	109	7	10	72.23	78.79	8.41	88.82	75.37
617	Mobility	35	102	10	9	79.55	77.78	8.11	87.83	78.66
618	Safety	30	125	7	7	81.09	81.09	5.31	91.72	81.09
619	Providers	24	121	17	8	75	58.54	6.21	85.3	65.76
620	Anxiety	62	137	8	12	83.79	88.58	8.06	90.87	86.12
621	Environmental	36	128	7	9	80	83.73	6.57	91.12	81.82
622	Emotional	40	138	8	7	85.11	83.34	4.83	92.23	84.22
623	Medical Issues	38	124	4	8	82.61	90.48	6.07	93.11	86.37
624	Insurance Issues	47	172	13	16	74.61	78.34	8.52	88.31	76.43
625	Medication Management	50	113	19	11	81.97	72.47	8.88	84.46	76.93

626	Activities Of Daily Living	37	116	6	10	78.73	86.05	7.94	90.54	82.23
627	Spiritual	27	114	9	3	90	75	2.57	92.16	81.82
628	Care Provision	21	146	7	12	63.64	75	7.6	89.79	68.86
629	Health History	53	99	6	8	86.89	89.84	7.48	91.57	88.34
630	Social	49	94	6	7	87.5	89.1	6.94	91.67	88.29
631	Well-Being	33	107	9	8	80.49	78.58	6.96	89.18	79.52
632	Supportive Services	36	122	12	8	81.82	75	6.16	88.77	78.27
633	Safety and Assistive Devices	22	114	10	5	81.49	68.75	4.21	90.07	74.58
634	Elder Abuse	24	82	7	10	70.59	77.42	10.87	86.18	73.85
635	Information Preference	22	145	6	11	66.67	78.58	7.06	90.77	72.14
636	Intellectual	36	87	7	7	83.73	83.73	7.45	89.79	83.73
637	Social Interaction	35	89	3	14	71.43	92.11	13.6	87.95	80.46
638	Preferences	37	132	9	8	82.23	80.44	5.72	90.87	81.32
639	Legal Older Adults	23	110	2	10	69.7	92	8.34	91.73	79.32
640	Medical Alerts	33	90	12	9	78.58	73.34	9.1	85.42	75.87
641	Chronic Disease	49	96	10	13	79.04	83.06	11.93	86.31	81

4.7 Comparison of Results

The results obtained from the experiments described and reported in section 4 have been compared with the results from the related work in Chapter 2. Table 4.11 shows the comparison. Some of the papers have reported only the overall accuracy of the methodologies. Also, some of the papers do not report the size of the data-set. Therefore, corresponding entries are not mentioned in the table.

From the comparison of the systems, it can be seen that the causal extraction system developed in this thesis is has comparable results. The precision and recall obtained shows the quality of the causal relations extracted by the system.

The list extracted in step 2 can contain duplicates which are removed in the final step.

Table 4.11: Performance Comparison of the System Described in this Thesis With Other Systems Discussed in Chapter 2

System	Method Adopted	Data-Set Used	Precision	Recall	Accuracy
Khoo et al.	Linguistic clues and Pattern-matching	Wall Street Journal - 1082 Sentences	25%.	68%.	Not Reported
Marcu et al.	Cue phrase filter	English texts and BLIPP Corpus	Not Reported	Not Reported	93%
Girju and Moldovan	Syntactic and semantic classification	TREC-9 Wall Street Journal	Not Reported	Not Reported	65%
Previous Work by us	Naive Bayes Classifier	Geriatric Data 60 Abstracts, 588 Sentences	35-40%	35-75%	68-78%
Previous Work by us	N-Gram Model	Geriatric Data 2280 Abstracts, 19725 Sentences	66%	74%	80%
System described in the Thesis	Semantic Tagging, Dictionary and CRF based	Geriatric Data 2280 Abstracts, 19725 Sentences	79.54%	81%	89%

5 CONCLUSION AND FUTURE WORK

5.1 Conclusion

Due to the complex nature of the literature, there is no one single method to achieve causal identification and it has to be done in a multi-layered form with different methods. Since the traditional classification and probabilistic methods did not yield satisfactory results, the use of a more in-depth semantic analysis and machine learning approach is the way to go for information extraction.

The work reported in this thesis shows a combination of semantic tagging, dictionaries and machine learning approaches. The causal relations are detected using the semantic tagging and the dictionaries and the actors in the relations are detected using the machine learning approach. Since the learning approach involved is feature based and not actual words, it gave accurate results over the different experiments. The confidence established by adopting this approach is also higher. The creation of the training data to learn the method is much easier which makes the system scalable.

The dictionaries used in the experiments have been constructed using the abstracts from the geriatric fields and to be more precise, from those selected from Pubmed. This aspect makes the dictionaries, a unfixed set. Adding to terms to the dictionary would have to be done in a systematic matter by POS tagging the newly found terms and the sentences and extracting the POS tag triplets and updating the dictionaries.

Since the system works in a multi-layered fashion, the results obtained have the best quality and avoid the incursion of any unwanted relations (noise). The system provides a precision of 79.54%, recall of 81% and an accuracy of 89% but the false positive rate of the system is at 8% which can be attributed to the misclassification of

some sentences during initial classification, or the fact that the POS tagger *Medpost* is 97% accurate which if improved, can increase the performance of the system.

5.2 Future Work

During the course of the project, we came across several other aspects of natural language.

- Sentence Co referencing: Many of the sentences have incomplete information. For example:

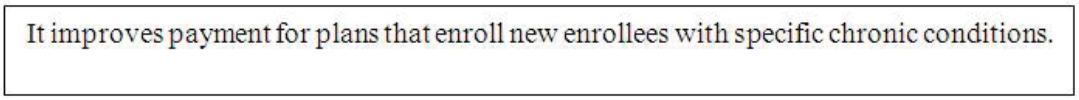

Figure 5.1.: Incomplete Sentence

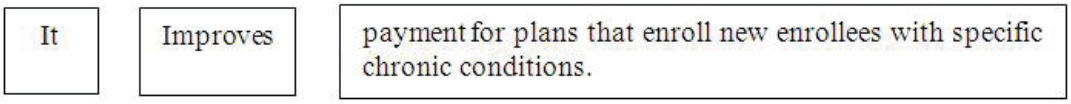

Figure 5.2.: Sentence Illustrating Coreferencing Issue

In Figure 5.2, the term "It" corresponds to an actor which can be referred to a different part of the abstract or text from which this sentence is taken. This needs co-referencing resolver which can provide suitable geriatric actors to become a part of the sentence before causal extraction can be applied. This step is shown in dotted blocks in the Figure 3.1.

Co-referencing can be analyzed by considering the different forms in which the sentences are organized. Figure 5.3, Figure 5.4 and Figure 5.5 shows few of the forms in which a sentence with causal content contains co-referencing issues.

Figure 5.3.: First Structure of Causal Sentence with Co-referencing

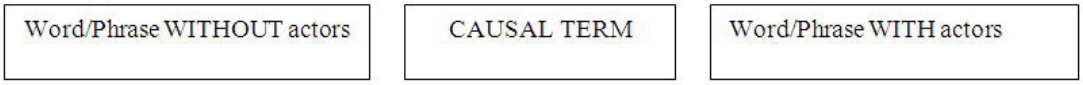

Figure 5.4.: Second Structure of Causal Sentence with Co-referencing

Figure 5.5.: Third Structure of Causal Sentence with Co-referencing

- Negation terms: Sentences that contain negation terms like *"no"*, *"none"*, *"not"* etc. were a part of the sentences that were analyzed. For example:

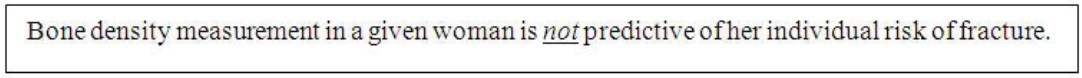

Figure 5.6.: Negated Sentence with "not"

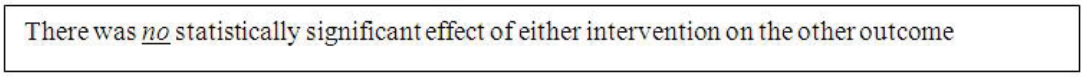

Figure 5.7.: Negated Sentence with "no"

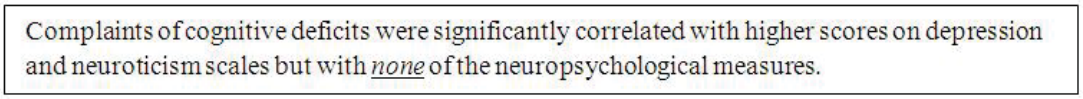

Figure 5.8.: Negated Sentence with "none"

The presence of these terms can also cause the causal extraction system to produce ambiguous results. The current system is capable of identifying these issues but has not been programmed to deal with them. Stanford NLP [43] [44], provides tools to that can handle co-referencing issues. Negex [45] provides a very good and efficient system that can be used to handle the Negation terms in the geriatric sentences.

This work can be used to build a more generic model that can address causal information extraction problems in various other fields. Apart from the geriatric domain, this can be applied on other biomedical fields to improve the performance of existing information extraction systems.

LIST OF REFERENCES

LIST OF REFERENCES

[1] Ah hwee Tan. Text mining: The state of the art and the challenges. In *In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, pages 65–70, 1999.

[2] Rogers M.L. Geriatric syndromes. 2008.

[3] H. Jiawei and M. Kamber. Data mining: concepts and techniques. *San Francisco, CA, itd: Morgan Kaufmann*, 5, 2001.

[4] S. Theodoridis and K. Koutroumbas. Pattern recognition. 2006.

[5] S.M. Weiss. *Text mining: predictive methods for analyzing unstructured information*. Springer-Verlag New York Inc, 2005.

[6] G. Dreyfus. *Neural networks: methodology and applications*. Springer-Verlag New York Inc, 2005.

[7] K. Gurney and K.N. Gurney. *An introduction to neural networks*. CRC Press, 1997.

[8] I. Steinwart and A. Christmann. *Support vector machines*. Springer Verlag, 2008.

[9] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Academic Pr, 1990.

[10] R. Feldman, Y. Regev, M. Finkelstein-Landau, E. Hurvitz, and B. Kogan. Mining biomedical literature using information extraction. *Current Drug Discovery*, 2(10):19–23, 2002.

[11] US National Library of Medicine National Institutes of Health.

[12] G.G. Chowdhury. Natural language processing. *Annual review of information science and technology*, 37(1):51–89, 2003.

[13] E.D. Liddy. Natural language processing. 2001.

[14] H. Van Halteren. *Syntactic wordclass tagging*, volume 9. Springer, 1999.

[15] E. Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational linguistics*, 21(4):543–565, 1995.

[16] K. Toutanova, D. Klein, C.D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

[17] L. Smith, T. Rindflesch, W.J. Wilbur, et al. Medpost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321, 2004.

[18] C.S.G. Khoo, J. Kornfilt, R.N. Oddy, and S.H. Myaeng. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, 13(4):177–186, 1998.

[19] D. Marcu and A. Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375. Association for Computational Linguistics, 2002.

[20] R. Girju and D. Moldovan. Text mining for causal relations. In *Proceedings of the FLAIRS Conference*, pages 360–364. AAAI Press, 2002.

[21] G.A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[22] C.F. Meyer. *English corpus linguistics: An introduction*. Cambridge Univ Pr, 2002.

[23] Y. Wilks and M. Stevenson. Sense tagging: Semantic tagging with a lexicon. In *Proceedings of the SIGLEX Workshop Tagging Text with Lexical Semantics: What, why and how*, pages 47–51, 1997.

[24] E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the workshop on Speech and Natural Language*, pages 112–116. Association for Computational Linguistics, 1992.

[25] S.D. Sudarsan. *Signal Detection Framework Using Semantic Text Mining Techniques*. ProQuest LLC. 789 East Eisenhower Parkway, PO Box 1346, Ann Arbor, MI 48106. Tel: 800-521-0600; Web site: http://www. proquest. com/en-US/products/dissertations/individuals. shtml, 2009.

[26] R. Durbin. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge Univ Pr, 1998.

[27] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 591–598, 2000.

[28] L.R. Rabiner and R.W. Schafer. Introduction to digital speech processing. *Foundations and trends in signal processing*, 1(1):1–194, 2007.

[29] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[30] N. Ye, W.S. Lee, H.L. Chieu, D. Wu, and S.M.I.T. Alliance. Conditional random fields with high-order features for sequence labeling. *Advances in Neural Information Processing Systems*, 22:2196–2204, 2009.

[31] C. Sutton and A. McCallum. *An introduction to conditional random fields for relational learning*. Introduction to statistical relational learning. MIT Press, 2006.

[32] A.K. McCallum. Mallet: A machine learning for language toolkit. 2002.

[33] J. Lafferty, A. McCallum, and F.C.N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[34] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics, 2005.

[35] E. Riloff and W. Phillips. An introduction to the sundance and autoslog systems. Technical report, Technical Report UUCS-04-015, School of Computing, University of Utah, 2004.

[36] R. Leaman, G. Gonzalez, et al. Banner: An executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, volume 13, pages 652–663, 2008.

[37] B. Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107. Association for Computational Linguistics, 2004.

[38] A. Culotta, A. McCallum, and J. Betz. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 296–303. Association for Computational Linguistics, 2006.

[39] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *In NIPS*. Citeseer, 2004.

[40] S.C. Suh. *Practical Applications of Data Mining*. Jones & Bartlett Publishers, 2011.

[41] L.P. Alias-I. 4.0. 0.

[42] I.H. Witten, E. Frank, and M.A. Hall. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

[43] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34. Association for Computational Linguistics, 2011.

[44] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics, 2010.

[45] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.