Applying Natural Language Processing to Clinical Information Retrieval

by

James Cogley B.A.

The thesis is submitted to University College Dublin in fulfilment of the requirements for the degree of Doctor of Philosophy

School of Computer Science and Informatics

Head of School: John Dunnion
Under the supervision of Prof. Joe Carthy and Dr. Nicola Stokes

February 2014
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4.1</td>
<td>Dataset Overview</td>
<td>54</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Systems for MER and Concept Normalization</td>
<td>56</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Evaluation Results</td>
<td>59</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Discussion</td>
<td>68</td>
</tr>
<tr>
<td>3.5</td>
<td>Summary</td>
<td>68</td>
</tr>
<tr>
<td>4</td>
<td>Assertion Classification in the Medical Domain: Experiments</td>
<td>71</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>71</td>
</tr>
<tr>
<td>4.2</td>
<td>Evaluation Metrics</td>
<td>72</td>
</tr>
<tr>
<td>4.3</td>
<td>UCD Condition Attribution and Temporal Grounding Corpus</td>
<td>73</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Experimental Dataset</td>
<td>73</td>
</tr>
<tr>
<td>4.3.2</td>
<td>System Descriptions</td>
<td>75</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Evaluation Results</td>
<td>79</td>
</tr>
<tr>
<td>4.4</td>
<td>Assertion Classification on the i2b2/VA dataset</td>
<td>87</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Dataset Overview</td>
<td>87</td>
</tr>
<tr>
<td>4.4.2</td>
<td>System Description</td>
<td>88</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Evaluation Results</td>
<td>90</td>
</tr>
<tr>
<td>4.5</td>
<td>Summary</td>
<td>96</td>
</tr>
<tr>
<td>5</td>
<td>Information Retrieval in the Medical World</td>
<td>99</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>99</td>
</tr>
<tr>
<td>5.2</td>
<td>Foundations of Information Retrieval</td>
<td>100</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Ranked Information Retrieval Models</td>
<td>101</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Evaluation of Information Retrieval</td>
<td>105</td>
</tr>
<tr>
<td>5.3</td>
<td>Information Retrieval in the Biomedical Domain</td>
<td>107</td>
</tr>
<tr>
<td>5.4</td>
<td>Information Retrieval in the Clinical Domain</td>
<td>110</td>
</tr>
<tr>
<td>5.4.1</td>
<td>TREC Medical Track</td>
<td>110</td>
</tr>
<tr>
<td>5.5</td>
<td>Summary</td>
<td>119</td>
</tr>
<tr>
<td>6</td>
<td>Implementing a Clinical IR approach</td>
<td>122</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>122</td>
</tr>
<tr>
<td>6.2</td>
<td>Experiment Dataset</td>
<td>123</td>
</tr>
<tr>
<td>6.3</td>
<td>Indri Retrieval System</td>
<td>125</td>
</tr>
<tr>
<td>6.4</td>
<td>Tools for retrieval</td>
<td>130</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Query Expansion Methods</td>
<td>130</td>
</tr>
<tr>
<td>6.5</td>
<td>Clinical IR Workflow</td>
<td>136</td>
</tr>
<tr>
<td>6.6</td>
<td>Summary</td>
<td>137</td>
</tr>
<tr>
<td>7</td>
<td>Experimental Evaluation of a Structured Clinical Information Retrieval Approach</td>
<td>139</td>
</tr>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>139</td>
</tr>
<tr>
<td>7.2</td>
<td>Experimental Setting</td>
<td>140</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Experimental Datasets</td>
<td>140</td>
</tr>
<tr>
<td>7.2.2</td>
<td>Experimental Evaluation</td>
<td>141</td>
</tr>
<tr>
<td>7.3</td>
<td>Applying Sentence-based MER to Queries</td>
<td>141</td>
</tr>
<tr>
<td>7.4</td>
<td>Experimentation on the TREC Medical Track 2011/2012 Dataset</td>
<td>143</td>
</tr>
<tr>
<td>7.4.1</td>
<td>Analysing the Impact of Annotating Problems, Treatments and Tests</td>
<td>143</td>
</tr>
</tbody>
</table>
7.4.2 Concept Normalisation and Query Expansion Coverage 149
7.4.3 Assertion Classification and Demographic Information 155
7.5 Summary .................................................. 160

8 Conclusions and Future Considerations ................................................................. 163
8.1 Introduction .................................................................................. 163
8.2 Thesis Contributions ........................................................................ 164
  8.2.1 Contributions to Natural Language Processing in the Clinical Domain ........................................................................ 164
  8.2.2 Contributions to Information Retrieval in the Clinical Domain ................................................................................ 166
8.3 Future Research Considerations ......................................................... 168
  8.3.1 State-of-the-art Concept Normalisation & Information Retrieval .................................................................................. 168
  8.3.2 Supervised Evidence Combination Methods ................................................. 169
  8.3.3 Novel Query Expansion & Weighting Normalised Concepts ............................................................... 170
  8.3.4 Improving Clinical NLP performance .......................................................... 171
8.4 Summary ........................................................................................... 173

A TREC Medical Track Query Set ................................................................................. 174

B Sample Report Documents .................................................................................. 177
  B.1 Radiology Report ................................................................................. 177
  B.2 Discharge Summary ............................................................................ 178
  B.3 History & Physical Examination Report ..................................................... 179
  B.4 Progress Notes Report ......................................................................... 180
  B.5 Surgical Pathology Report ...................................................................... 181
  B.6 Operative Report ................................................................................ 182
  B.7 Emergency Room Report ...................................................................... 183
  B.8 Echocardiogram Report ....................................................................... 185
LIST OF TABLES

2.1 List of i2b2 Shared Tasks 2008-2011 .................................. 24
2.2 Top Concept Identification Systems at i2b2 2010: Machine Learning Algorithms ................................................. 27
2.3 Top Concept Identification Systems at i2b2 2010: Feature-Sets ............................................................... 28
2.4 Top Assertion Classification Systems at i2b2 2010: Machine Learning Algorithms ................................................. 30
2.5 Top Assertion Classification Systems at i2b2 2010: Feature-Sets ............................................................... 31
3.1 Calculating TP, FP and FN for Machine Learning Evaluation ............................................................... 39
3.2 Examples of Problem, Treatment and Test concepts in the i2b2 Challenge Corpus .................................................... 41
3.3 Annotated Concept Counts in i2b2 Challenge Corpus . . . . . . 42
3.4 Lexical Features used in UCD_MER System ....................... 45
3.5 UCD_MER: BIO Vs. BIESO Tagging on the i2b2 Corpus . . . 48
3.6 Exact and Inexact evaluation of UCD_MER using all features ............................................................... 49
3.7 Breakdown of UCD_MER performance per entity type . . . . . 49
3.8 Inexact evaluation of UCD_MER, ABNER and Stanford NER on i2b2 Challenge Corpus .................................................... 52
3.9 Breakdown of CLEF/ShARe Concept Annotations .................. 55
3.10 Official Submission Results for CLEF/ShARe Task 1a, Exact Evaluation ............................................................... 59
3.11 Official Submission Results for CLEF/ShARe Task 1a, Inexact Evaluation ............................................................... 60
3.12 Disjoint and Joint F-scores for author’s submissions CLEF/ShARe Task 1a, Exact Evaluation ................... 60
3.13 Top 10 Performing Systems at CLEF/ShARe Task 1a, Exact Evaluation ............................................................... 61
3.14 Top 10 Performing Systems at CLEF/ShARe Task 1a, Inexact Evaluation ............................................................... 61
3.15 Feature Selection for MER for CLEF/ShARe task, F-score Results. ............................................................... 64
3.16 Official Author Submission Results for CLEF/ShARe Task 1b: Concept Normalisation .................................................... 66
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.17</td>
<td>Top 10 Performing Systems for CLEF/ShARe Task 1b, Exact Evaluation</td>
<td>66</td>
</tr>
<tr>
<td>3.18</td>
<td>Top 10 Performing Systems for CLEF/ShARe Task 1b, Inexact Evaluation</td>
<td>67</td>
</tr>
<tr>
<td>3.19</td>
<td>Effects of Feature Selection on Concept Normalisation</td>
<td>67</td>
</tr>
<tr>
<td>3.20</td>
<td>Effects of Edit-Distance CUI Selection on Concept Normalisation</td>
<td>67</td>
</tr>
<tr>
<td>4.1</td>
<td>Condition Attribution Instances in the UCD Condition Attribution and Temporal Grounding Corpus</td>
<td>74</td>
</tr>
<tr>
<td>4.2</td>
<td>Temporal Grounding Instances in the UCD Condition Attribution and Temporal Grounding Corpus</td>
<td>74</td>
</tr>
<tr>
<td>4.3</td>
<td>Trigger List Feature Descriptions for Binary Assertion Classification</td>
<td>77</td>
</tr>
<tr>
<td>4.4</td>
<td>Contextual Feature Descriptions for Binary Assertion Classification</td>
<td>78</td>
</tr>
<tr>
<td>4.5</td>
<td>Syntactic Feature for Binary Assertion Classification</td>
<td>79</td>
</tr>
<tr>
<td>4.6</td>
<td>UCD Temporal Grounding &amp; Condition Attribution Feature-Set Combinations</td>
<td>81</td>
</tr>
<tr>
<td>4.7</td>
<td>Condition Attribution: Cross-Validation Results of Classification Algorithms on UCD Dataset</td>
<td>81</td>
</tr>
<tr>
<td>4.8</td>
<td>Condition Attribution: Performance Results on 30% Training Data Vs. ConText on UCD Dataset</td>
<td>82</td>
</tr>
<tr>
<td>4.9</td>
<td>Temporal Grounding Cross-Validation Results of Classification Algorithms on UCD Dataset</td>
<td>84</td>
</tr>
<tr>
<td>4.10</td>
<td>Temporal Grounding Performance Results on 30% Training Data Vs. ConText on UCD Dataset</td>
<td>84</td>
</tr>
<tr>
<td>4.11</td>
<td>Examples of Assertion Status in the i2b2 Challenge Corpus</td>
<td>88</td>
</tr>
<tr>
<td>4.12</td>
<td>Assertion Status Counts in i2b2 Challenge Corpus</td>
<td>88</td>
</tr>
<tr>
<td>4.13</td>
<td>Results of extending Binary feature-set to Multi-Class Assertion Classification Using Random Forest</td>
<td>89</td>
</tr>
<tr>
<td>4.14</td>
<td>Features added to Multi-class Assertion Classification approach</td>
<td>90</td>
</tr>
<tr>
<td>4.15</td>
<td>OVO Vs. OVA Multi-class Assertion Classification using LibSVM Linear Kernel</td>
<td>91</td>
</tr>
<tr>
<td>4.16</td>
<td>Kernel Selection for Assertion Classification: RBF Vs. Linear Kernel</td>
<td>91</td>
</tr>
<tr>
<td>4.17</td>
<td>Top performing feature sets for i2b2 Assertion Classification task, Micro and Macro Average</td>
<td>93</td>
</tr>
<tr>
<td>4.18</td>
<td>Top performing feature sets for i2b2 Assertion Classification, class breakdown</td>
<td>93</td>
</tr>
<tr>
<td>5.1</td>
<td>A Sample Inverted Index</td>
<td>101</td>
</tr>
<tr>
<td>5.2</td>
<td>A Sample Inverted Index With Term Frequency Information</td>
<td>101</td>
</tr>
<tr>
<td>5.3</td>
<td>Approaches used at TREC Medical Track 2011</td>
<td>115</td>
</tr>
<tr>
<td>5.4</td>
<td>Approaches used at TREC Medical Track 2012</td>
<td>118</td>
</tr>
<tr>
<td>6.1</td>
<td>Initial Allowed Relations between Concepts in the UMLS</td>
<td>132</td>
</tr>
</tbody>
</table>
7.1 Applying MER to TREC 2011 and 2012 Query Sets: Evaluation 142
7.2 BPref scores of IR systems leveraging MER on TREC 2011 & 2012 query sets. .................................................. 144
7.3 BPref scores of combined IR systems using MER on TREC 2011 & 2012 query sets with Relaxed Windows ............ 146
7.4 Effects of Expansion Resource Choice on Query Length .... 149
7.5 BPref Performance of Expansion Strategies on TREC 2011 & 2012 query sets. .................................................. 150
7.6 BPref Performance of Normalisation Strategies on TREC 2011 & 2012 query sets. .................................................. 154
7.7 BPref Performance Concept Normalisation Approaches with Age and Gender filtering. ................................. 157
7.8 BPref results of Applying Assertion Classification to Concept Normalisation. .................................................. 158
A.1 Topics from the 2011 TREC Medical Track ........................ 175
A.2 Topics from the 2012 TREC Medical Track ........................ 176
# LIST OF FIGURES

1.1 EHR Adoption among US Family Physicians .......................... 2
1.2 Examples of Natural Language Clinical Queries ....................... 3
1.3 Example Electronic Health Record ................................. 4
1.4 NLP and IR integration workflow ................................... 4

2.1 Pipeline for NLP for Biomedical and Clinical Domains ............ 11
2.2 Example Dependency Parse Tree .................................. 14
2.3 Example Binary Classification Problem .............................. 15

3.1 Diagram of Chapter 3 Systems and Datasets ....................... 39
3.2 Example of BIO tagging format .................................... 43
3.3 Example of Hierarchical Clustering ................................ 47
3.4 Workflow Overview of UCD_MER System .......................... 48
3.5 UCD_MER Versus Stanford NER and ABNER (Exact Evaluation) .......................................................... 50
3.6 UCD_MER Feature Performance on i2b2 Challenge Subset ....... 52
3.7 Overview of Disorder Recognition and Concept Normalization system using SSVM ............................................. 56
3.8 Accounting for tokens between disjoint concepts .................... 56
3.9 Example of BIESTO tagging ....................................... 56
3.10 Exact Evaluation for Feature Sets for CLEF/ShARe Task 1a ....... 63
3.11 Inexact Evaluation for Feature Sets for CLEF/ShARe Task 1a ...... 63

4.1 Diagram of Chapter 4 Systems and Datasets ....................... 72
4.2 Feature-Set Precision Performance for Condition Attribution ...... 82
4.3 Feature-Set Precision Performance for Temporal Grounding ...... 85
4.4 Feature-Set Recall Performance for Temporal Grounding ........ 86
4.5 Feature-Set F-score Performance for Temporal Grounding ....... 86
4.6 Examples of the class Possible in the i2b2 Assertion Classification task .......................................................... 89
4.7 Example of Absent misclassified as Present ....................... 95
4.8 Example of Possible Misclassified as Present ....................... 95
4.9 Example of Present misclassified as Absent ....................... 96

6.1 Diagram of Information Retrieval System and Datasets .......... 123
ABSTRACT

Medical literature, such as medical health records are increasingly digitised. As with any large growth of digital data, methods must be developed to manage data as well as to extract any important information. Information Retrieval (IR) techniques, for instance search engines, provide an intuitive medium in locating important information among large volumes of data. With more and more patient records being digitised, the use of search engines in a healthcare setting provides a highly promising method for efficiently overcoming the problem of information overload.

Traditional IR approaches often perform retrieval based solely using term frequency counts, known as a ‘bag-of-words’ approach. While these approaches are effective in certain settings they fail to account for more complex semantic relationships that are often more prevalent in medical literature such as negation (e.g. ‘absence of palpitations’), temporality (e.g. ‘previous admission for fracture’) or attribution (e.g. ‘Father is diabetic’), or even term dependencies (“colon cancer”). Furthermore, the high level of linguistic variation and synonymy found in clinical reports gives rise to issues of vocabulary mismatch whereby concepts in a document and query may be the same, however given differences in their textual representation relevant documents are missed e.g. hypertension and HNT. Given the high cost associated with errors in the medical domain, precise retrieval and reduction of errors is imperative.

Given the growing number of shared tasks in the domain of Clinical Natural Language Processing (NLP), this thesis investigates how best to integrate Clinical NLP technologies into a Clinical Information Retrieval workflow in order to enhance the search engine experience of healthcare professionals. To determine this we apply three current directions in Clinical NLP research to the retrieval task. First, we integrate a Medical Entity Recognition system, developed and evaluated on I2B2 datasets, achieving an f-score of 0.85. The second technique clarifies the Assertion Status of medical conditions by determining who is the actual experiencer of the medical condition in the report, its negation and its temporality. Standalone evaluations on I2B2 datasets, have seen the system achieve a micro f-score of 0.91. The final NLP technique applied is that of Concept Normalisation, whereby textual concepts are mapped to concepts in an ontology in order to avoid problems of vocabulary mismatch. While evaluation scores on the CLEF evaluation corpus are 0.509, this concept normalisation approach is shown in the thesis to be the most effective NLP approach of the three explored in aiding Clinical IR performance.
I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at this, or any other, University or institute of tertiary education.

James Cogley
February 2014
ACKNOWLEDGEMENTS

First and foremost, I would like to thank Joe and Nicola for taking me on board, without that one decision none of this would be possible. Nicola, thank you for your friendship, unending support and dedication throughout my time at UCD. You’ve always helped me look on the bright side, kept me thinking positively and critically to see things through.

To my parents, Susan and Aidan, you’re the ones that started it all. Even though you may not understand what I go on about half the time, you’ve dedicated everything to get me to university, and encouraged me to go further.

To Mikey, you’ve always had an interest in anything I’ve done and have always kept me motivated. I also blame you for bringing out the math nerd in me while helping me with my “sums” throughout the years.

Thanks JP, Willem and the rest of the Microsoft guys. You all injected much needed focus and direction to my work as well as made my time with you as enjoyable as possible.

To Chris, Mark and Rebecca, you guys provided the much needed breaks from the plethora of code, reading and writing and I’m incredibly fortunate to have you all around.

To Aonghus, our conversations have ranged from the academic to the absurd, if there was ever a complacent part of me, you’ve rid me of it. Thanks for the proofing, the discussions and just being there. Maybe now I can get a head start on those books.

Thanks to Dave, a great roommate and an even better friend. You’ve been there to console me when things went wrong, and to celebrate them when they go right.

To Emma, you’ve had to listen to (or suffer?) all the convoluted explanations and hand waving more than anyone else, but you’ve still stuck by me. I’ll repay your efforts this year, but for now, I am incredibly grateful thank you for your love and support that has kept me pushing to reach my goals. I promise I’ll shave now.


J. Cogley, N. Stokes, J. Carthy and J. Dunnion. *UCD IIRG at TREC 2012 Medical Track,* Proceedings of the 21\textsuperscript{st} Text Retrieval Conference, 2012, Gaithersburg, United States

J. Cogley, N. Stokes, J. Carthy and J. Dunnion. *UCD IIRG at TREC 2011 Medical Track,* Proceedings of the 20\textsuperscript{th} Text Retrieval Conference, 2011, Gaithersburg, United States

J. Cogley, *A Linguistically Enhanced Search Engine For Identifying Suitable Patients For Clinical Trials,* The 16\textsuperscript{th} HISI Annual Conference, 2011
Healthcare is constantly striving for the total eradication of errors in its workflow. While the health and well-being of patients is paramount, errors in initial care also generate unnecessary costs in returning visits for patients. One root of errors made in healthcare is the fragmented nature of paper-based patient records, which has contributed to 98,000 annual deaths and the cost of over 29 billion dollars annually in the United States alone (Cohn et al., 1999). Given the clear threat to patient’s health and cost-factors, governments around the world have focused on bringing healthcare into the digital age. Policies such as the Health Information Technology for Economic and Clinical Health (HITECH) Act ¹ and Directive 2011/24/EU ² have incentivised the use of Electronic Health Records, a digital replacement for the recording of a patient’s health. Though initially expensive to set up, EHR systems increase the efficiency of workflow while reducing the rate of errors showing that the benefits vastly outweigh initial costs (Hillestad et al., 2005). These benefits and intervention of government directives have seen the adoption of EHR grow steadily over the past ten years as indicated in Figure 1.1. The growth in adoption of this new form of data brings with it a set of challenges and opportunities in providing state-of-the-art next-generation healthcare.

Up until the turn of the millennium, less than 20% of healthcare facilities had in-house EHR facilities. A number which is constantly climbing. The integration of EHR gives healthcare institutions the ability to mine data in order to provide superior healthcare. EHRs have a great potential in clinical-decision support systems that provide clinicians a monitoring system that gives timely reminders and recommendations (Hoffman, 2009) or retrieves patients who

¹http://www.healthit.gov/policy-researchers-implementers/hitech-act-0
are eligible for clinical trials. Far more ambitious projects may be undertaken in order to pre-emptively identify disease outbreaks (Greene et al., 2012) or perform predictive diagnosis of patients based on previous admissions (McCormick et al., 2012).

To facilitate such tasks, systems must first be able to effectively and efficiently manage and retrieve Electronic Health Records. There are two key stipulations that determine the successful adoption of such systems: First, the impact and requirement of training of healthcare professionals should be minimal as too many restrictions may hinder their overall efficacy due to drastic changes in the workflow (Hoffman, 2009). Secondly, systems must handle unstructured textual information. While there are structured resources available, such as in the MEDLINE database\(^3\), it is reported that over 80% of data available to clinicians is unstructured (Hale, 2005), as shown in Figure 1.3. These stipulations rule-out the use of structured relational database systems due to the requirement of structured information as well as knowledge of a structured query language.

This leads to the investigation of the usage of Information Retrieval (IR) methods that allow users to identify documents in a collection that are relevant to their information need. An information need is typically expressed by a user in the form of a textual query. Example queries are shown in Figure 1.2.

The criteria for determining a document is relevant to a query typically uses keyword search. Keyword search deems all documents which contain tokens in

---

\(^3\)www.ncbi.nih.gov/pubmed
a given query relevant to that query.

Keyword search is a baseline method for Information Retrieval, with a similarity measure generated between document and query indicating the level of relevance for a document to a given query. However, this has many shortcomings, especially in the clinical domain. For example in applying Query 1 from Figure 1.2 we would incorrectly retrieve the document in Figure 1.3 from the document collection as it matches the terms cancer and vomiting. In this example, keyword querying lacks essential term dependency information as the query is satisfied by mentions of cancer rather than the correct breast cancer. Query 2 will also incorrectly retrieve the example document as both have mentions of alzheimer’s and heart failure. However, on inspection of the document we see the respective conditions affect the patient’s relatives, rather than the patient himself. In this case there is a mismatch in the assertion status of the medical conditions in the query and those found in the document. Query 3 exhibits the vocabulary mismatch problem. In this case the document is relevant to the query, as CHF is semantically equivalent to heart failure, however as different terms are used to express the same concept, the document is deemed irrelevant.

| 1. Patients with breast cancer exhibiting vomiting from therapy. |
| 2. Patients with alzheimer’s admitted with heart failure. |
| 3. Patients with family history of CHF. |

Figure 1.2: Examples of Natural Language Clinical Queries

In the course of this thesis, we analyse the impact of applying Natural Language Processing techniques to Clinical Information Retrieval to alleviate the issues described above. However, unlike previous approaches we examine the contribution of several techniques from multiple shared tasks in the clinical NLP domain before applying them to a Clinical Information Retrieval workflow. In particular, we analyse the contribution of three Clinical NLP methodologies: Medical Entity Recognition, Concept Normalisation and Assertion Classification.

Medical Entity Recognition is an extension of Named Entity Recognition in the general domain. Rather than identify company or location names in a text, MER aims to identify medical concepts in the text such as problems experienced by the patient, treatments and tests. This system in an IR context facilitates the
HISTORY OF PRESENT ILLNESS:
Caucasian male, **AGE[in 40s]**, with resolved colon cancer presented with migraines and vomiting. He worked as a mechanic and may have had some mineral asbestos exposure with brake linings.

ALLERGIES
NO KNOWN ALLERGIES.

FAMILY HISTORY:
Aunt known to have alzheimer’s.
Father died at 54 due to heart failure.

PAST MEDICAL HISTORY:
The patient has a history of diabetes and hypertension with resolved colon cancer.

PHYSICAL EXAMINATION:
On examination, lying in bed, awake and alert, in no apparent distress.
HEENT examination: He had anicteric sclerae. Moist mucous membranes. He had poor dentition.
Cardiovascular: Regular rate and rhythm, no murmurs, no lower extremity edema.

ASSESSMENT AND PLAN
Mr. **NAME[AAA]** is an **AGE[in 40s]**-year-old pack per day smoker with chronic migraines and vomiting. He is on Coumadin and has previously been taking aspirin 3 times daily.
I will discontinue aspirin, and prescribe imitrex, which should reduce the patients migraines and nausea.

---

Figure 1.3: Example Electronic Health Record

Figure 1.4: NLP and IR integration workflow

automatic identification of term dependencies in queries. Another increasingly popular Clinical NLP task is Assertion Classification, in which systems assign an assertion status to concepts in the text. In its most basic form assertion classification determines if a medical condition is affecting a patient or not.

Concept Normalisation extracts concepts from free text and normalises them with respect to a gold standard ontology to alleviate issues of vocabulary mismatch. Further to the application of NLP approaches, we investigate the application of demographic information in order to identify cohorts among the document collection as well as determining the impact of query expansion.
The resulting Clinical Information Retrieval system requires a complex structured query language to leverage linguistic information generated by the NLP processes, as shown in Figure 1.4. To remove the need for a user’s expertise in the structured query language, the user inputs queries as given in Figure 1.2 that are then translated into a structured format by the IR system. These three NLP technologies logically follow one another. For example, to perform concept normalisation entities must first be identified using the Medical Entity Recognition system. This is similarly the case in using assertion classification, we must first identify entities.

The use of shared task datasets allows us to analyse the performance of several algorithms applied to the tasks, as well as exploratory feature analysis. Through applying these NLP tasks, we may determine the validity of these shared tasks and identify NLP approaches that alleviate issues in Clinical Information Retrieval.

## 1.1 Objectives

This thesis aims to investigate the impact of applying Natural Language Processing (NLP) techniques to Information Retrieval (IR) systems in the medical domain. In order to comprehensively address this question we set out the following objectives.

### Clinical NLP Goals

- How do rule-based systems compare to machine learning performance in Assertion Classification? Initial experimentation in the area began with rule-based methods borne out of strict linguistic analysis, with machine learning approaches implemented in more recent systems due to their adaptability. This thesis investigates the applicability of these approaches.

- What machine learning approaches best aid the tasks of Medical Entity Recognition and Assertion Classification? Given the vast array of machine learning approaches available, we examine what algorithms achieve top performance.

- What impact do different tagging formats have on MER performance? While systems typically use the traditional BIO format, further state-of-
the-art approaches have implemented the BIESO format. We investigate the impact and advantages of both approaches.

- How should a traditional MER system be adapted for the recognition of disjoint entities? Research in the domain of MER has moved towards the recognition of entities that may be divided by non-concept tokens. We investigate the problems faced in adopting a traditional MER system to this new challenge.

- What is the optimal feature space for machine learning based approaches? Often machine learning approaches contain thousands of features when performing these tasks. However, such expansive feature spaces are computationally expensive. In a healthcare setting, where centralised storage of data may not be possible, needless computation with noisy features should be avoided. In this thesis, we perform thorough feature selection to identify features that make the most difference in terms of performance.

Clinical Information Retrieval Goals

- Clinical Natural Language Processing techniques, such as Concept Identification and Assertion Classification are often cited in shared tasks as important application areas to next generation Information Retrieval approaches. We ask, what NLP tasks contribute most to Information Retrieval?

- What is the required performance of NLP, in order to aid IR approaches? Often investigations into applying NLP to IR that feature one or more technologies analyse their sum, rather than the contribution of their constituent components. In this work, IR experiments are performed at each stage of applying an NLP approach to the framework. Performing analysis in this manner will allow us to determine which NLP techniques aid or hinder retrieval effectiveness.

- What are the challenges faced in using NLP to automatically translate natural language queries to a structured query language? NLP techniques, such as Medical Entity Recognition are typically performed on sentences found in clinical reports. However, for its application to the Information Retrieval domain these systems must be applied to natural language queries. We investigate the exact challenges in applying NLP to queries.
What challenges are faced in performing query expansion in the clinical domain? Multiple resources and ontologies are available in the clinical domain to perform query expansion. We investigate the advantages and disadvantages of using two such resources, the MeSH ontology and the UMLS Metathesaurus.

To create an Information Retrieval system that leverages structured information generated from NLP system. Clinical texts are composed of free text fields, exhibiting nuances of natural language that prove difficult to understand using standard keyword queries. While a structured approach requires a significant human effort, we investigate how NLP techniques may automatically add structure to queries and documents thus providing the high-precision framework required of the medical domain.

1.2 Thesis Organisation

In seeking answers to these research questions, the dissertation is structured into two parts. First we investigate Natural Language Processing in the Clinical domain from Chapter 2 through Chapter 4. The second component, chapters 5 to 7, analyses the combination of Natural Language Processing techniques and an Information Retrieval framework in the clinical domain. In more detail, the chapters are broken down into the following areas of focus:

Chapter 2 provides a reader for Natural Language Processing techniques in the biomedical and clinical domains. In particular we discuss the theoretical background of NLP and machine learning techniques, as well as their evaluation. Moving on from a discussion of core concepts we analyse previous work among the community in the tasks of Medical Entity Recognition and Assertion Classification, particularly in reference to the seminal shared tasks such as the i2b2/VA challenges. This chapter provides the basis of our experimentation in future chapters.

Chapter 3 describes the development and experimental evaluations of Medical Entity Recognition and Concept Normalisation systems in the clinical domain. The systems developed build on previous work through extensive feature contribution analysis. Evaluations are performed using metrics and datasets as part of previous and current shared tasks in the domain.

Chapter 4 provides an overview of the development and evaluation of an
Assertion Classification system, first for binary classification followed by an extension to a multi-class problem. Experiments in Assertion Classification identify the Machine Learning approach best suited to the task as well as an extensive feature analysis.

Chapter 5 begins the second component of this thesis in which we investigate the application of the NLP techniques described in Chapter 3 and Chapter 4 to an Information Retrieval workflow. This chapter provides the theoretical background to Information Retrieval in the clinical and biomedical domains along with a discussion of previous methods in the domain which provide the basis of this thesis.

Chapter 6 describes the workflow of a Clinical Information Retrieval system that leverages the outputs of NLP methods. Further, the chapter presents the datasets and metrics used in evaluating the Information Retrieval approach.

Chapter 7 presents the experiments in Clinical Information Retrieval, allowing us to measure the impact of incorporating structured information, generated by NLP systems, into an Information Retrieval workflow. The chapter presents the evaluations of systems of two datasets available to the research community.

Chapter 8 concludes the dissertation. It provides a discussion of the goals of the thesis and measures the outcomes of experimentation toward them. A breakdown of contributions provided by this thesis to the NLP and IR communities in the clinical domain is provided and used in our reasoning for future considerations for work in the area.
2.1 Introduction

The advent of open-access journals combined with ubiquitous web accessibility has seen biomedical literature experience exponential growth (Cohen and Hersh, 2005). This poses a problem for biomedical researchers who must filter through this huge dataset in order to keep up to date with state-of-the-art research. In a clinical setting, we also see a great amount of data generated in the form of electronic health records. During a patient’s visit to a healthcare facility, a physician may create a multitude of such reports, for example emergency room admissions, radiology reports, lab results and discharge summaries. These reports create issues in data management and storage for ease of retrieval. However, this constant build-up of data also provides opportunities in developing novel methods to aid the healthcare process.

While these respective datasets contain highly important and interesting data, it is expressed in unstructured free texts. This opens the field to NLP researchers who aim to add structure to texts, mine and extract pertinent information and present it to the user in a meaningful and concise manner.

**What is biomedical text processing?**

In the biomedical domain, the literature is concerned with biological facts. Papers detail protein interactions, gene discoveries and relationships between entities, all through experimental methods. NLP applications strive to extract
this information in an accurate and concise manner. Applications may range from the summarisation of findings or experimental methodology (Shang et al., 2011) to protein-protein interaction discovery (MadKour et al., 2007), identifying relationships between entities (Kim et al., 2011a) leading to more ambitious projects such as protein structure identification.

What is clinical text processing?

The growth of data in the clinical domain has also attracted interest among the NLP community; however, developments in the area have been stifled by difficulties in data sharing, the lack of annotated datasets and shared standards creating difficulties in reproducing experiments (Chapman et al., 2011). Despite these obstacles NLP shows promise in creating useful applications in the clinical domain. Similar to the biomedical domain, clinical reports contain an ever growing amount of important information that is expressed in free text. However, unlike biomedical data such as Medline abstracts, clinical data is comprised of records of patients healthcare in a healthcare institution. Though initial research in the area focused on fact extraction and the structuring of clinical data (Anderson et al., 1975), more recent work has focused on novel usage of this data source for the benefit of healthcare (Feblowitz et al., 2011). Possible applications include clinical decision support systems that may further inform physician’s decision making (Demner-Fushman et al., 2009) or fully automated outbreak monitoring systems that process clinical reports and may pre-emptively signal the outbreak of a disease among a community (Chapman et al., 2004).

Though the specific aims of NLP applications in clinical and biomedical research domains may differ, they both provide many exciting opportunities for automatic text processing and information techniques to be applied with the overall goal of improving the quality, both for healthcare professionals and patients. NLP applications are pipelined as in Figure 2.1 in both clinical and biomedical domains, where lower level pre-processing of text facilitates higher level linguistic analysis and annotation.

In this chapter we present a discussion of NLP methods in the biomedical and clinical domains that form the pipeline in creating the required outputs. In Chapter 3 and Chapter 4, we will build on our discussion in this chapter, implementing and evaluating aspects of the NLP pipeline before it is applied to an Information Retrieval system in the clinical domain. The aims of this the chapter are set out as follows:
Figure 2.1: Pipeline for NLP for Biomedical and Clinical Domains

- What are the tasks of Medical Entity Recognition and Assertion Classification?
- What is the theoretical background to approaches in Natural Language Processing in the Clinical and Biomedical domains?
- What machine learning approaches perform best in Medical Entity Recognition?
- What features best aid the task of Medical Entity Recognition?
- What machine learning approaches perform best in Assertion Classification?
- What features best aid the task of Assertion Classification?

In Section 2.2 we provide a succinct yet comprehensive overview of Natural Language Processing techniques, in particular in reference to the biomedical and clinical domains. Sections 2.3 and 2.4 provide an overview of NLP research in the biomedical and clinical domains, respectively. From this point on we focus on clinical NLP, discussing the i2b2 shared tasks in Section 2.5 and CLEF/ShARe shared task in Section 2.6. Finally, in Section 2.7, we will define the NLP tasks investigated in relevance to the Information Retrieval domain.
2.2 Foundations of NLP

This section presents an overview of the theoretical background underpinning the approaches and techniques in NLP, as well as available toolkits which facilitate research in the domain. The range of applications of NLP in the biomedical and clinical domains are expansive. As such, the goal of this section is to provide the reader with an understanding of NLP workflow in both domains, with descriptions of terminology and a discussion of issues in the biomedical and clinical domains. For more in-depth coverage of NLP in the domain, the reader is encouraged to find these approaches in Ananiadou and McNaught (2006).

2.2.1 Fundamentals of Text Processing

In this section, we present key concepts and tasks which are fundamental to performing any work on natural language texts.

2.2.1.1 Pre-Processing a Clinical Corpus

As shown in Figure 2.1, to perform any meaningful NLP task on a document collection it must first undergo Pre-Processing. Pre-Processing a corpus may involve sentence splitting which places each sentence in a document to a new line. This normalisation of the corpus aids further tasks in the pipeline that require sentential information. Following sentence splitting we may perform the next pre-processing step tokenisation. Tokenisation splits a sentence into its constituent tokens. The analysis on a token-by-token basis again facilitates further tasks. In the next three sections, we will investigate methods of Linguistic Analysis which build on the Pre-Processing stage.

2.2.1.2 Lexical Normalisation & Morphological Analysis

Tokens in clinical texts may vary in their surface form due to inflections e.g. plurality, while still conveying the same core semantic information, such as the presence of a medical condition. To generalise information relating to tokens, we may normalise tokens through stemming and lemmatisation. Stemming algorithms such as the Porter algorithm (Porter, 1980) reduce word forms to a stem, eliminating any inflectional information, e.g. effusions would become ef-
fusion. Stemming is in essence the removal of a suffix. Lemmatisation is a more linguistically detailed method that analyses the part-of-speech of a token in order to discover its root and derivations.

In some cases, affix information may be beneficial to analysis as it allows the generalisation of lexical forms. For example, prefixes may indicate locations e.g., cardio-, trach- while suffixes may indicate the performance of a procedure e.g. -tomy. This generalisation ability of stemming will be leveraged in this work, rather than lemmatisation.

2.2.1.3 Part-Of-Speech Tagging

A part-of-speech (POS) tag indicates the lexical category to which a word form belongs, e.g., noun, verb etc. A POS tagger is an automated system that, given a collection of word forms, typically in a sentence and assigns a POS tag to each word form. POS-tagging is a non-trivial task as a word form’s POS tag is dependent on its surrounding context. For example, pain in arm may assign arm to be a noun; however arm pain may assign arm to be an adjective. Given its non-trivial nature, POS-tagging systems often utilise machine learning techniques, in particular Markov processes that leverage structural information in performing sequence labelling. This will be further described in Section 2.2.2.

2.2.1.4 Dependency Tree Parsing

Dependency parsing builds on the previous task of POS tagging by building a tree like structure to analyse the syntax of a sentence, such as that shown in Figure 2.2. At the leaf nodes are the word forms, with POS tags as parent nodes. At each level of the tree structure, nodes are linked into larger and larger constituent chunks or phrases. The top-level node is S, indicating the sentence itself. Links between constituents are dependency relations. The resolution of these relations between tokens allow us to reason about further NLP tasks of Linguistic Annotation such as Named Entity Recognition or Assertion Classification. Investigations and experimentation in the application of these parsing methods to Assertion Classification in particular are described in Section 4.3.2.2.

2.2.2 Machine Learning and NLP

Machine Learning tools build models of the world around them through a variety of methods in order to perform tasks. Machine Learning is applied
often to tasks where there is a high level of variability, that cannot be accounted for using simpler rule-based algorithms. Given the high level of variation available in natural language, machine learning techniques are often the preferred choice in developing NLP applications.

There are three key schools of thought in machine learning, namely supervised, unsupervised and semi-supervised methods. Supervised methods require labelled training data. Supervised systems use the labelled training dataset to learn a model that may then be applied to the task in order to apply labels to instances in the test set. For unsupervised methods, the training data presented to the approach is unlabelled. While supervised methods typically perform classification, unsupervised methods are utilised to discover unknown information in the dataset. For example, clustering is an unsupervised machine learning technique used to group similar data points in the dataset. Between these two methodologies lie semi-supervised approaches. Semi-supervised approaches utilise both unlabelled and labelled training data in order address shortcomings in the availability of labelled data, which requires manual expert annotation.

For the purposes of this thesis, we investigate both supervised and unsupervised methods. While the machine learning applications in this thesis are primarily supervised, techniques such as clustering are a purely unsupervised approach to machine learning that are used later in the work to aid feature creation.
2.2.2.1 The Task of Classification

The task of classification may be defined as follows: Given a sample space, \( D \) and a set of possible class labels, \( C \), it is the role of the classifier to assign one of the class labels of \( C \) to each instance in \( D \). For example, in Figure 2.3 we see a visualisation of a binary classification problem. An instance may be either positive or negative in this case, with the role of the classifier to correctly assign a new instance to the correct side i.e. class, of the boundary. Each instance in the sample space is represented as a feature vector. A feature vector expresses further information about the instance that may aid the classifier in generalising its task, e.g. for a named entity task, a POS tag may be a feature for tokens. Allowed features may include boolean, real or nominal based information, depending on the classifier used.

![Figure 2.3: Example Binary Classification Problem](image)

Machine Learning based classification algorithms range in their theoretical background and application areas. To discover machine learning algorithms best-suited to the NLP tasks in this thesis, we will investigate five classification algorithms. These classifiers have been chosen to represent different schools of thought in machine learning based classification. Each has its points for and against, in this work we will investigate how each approach functions on medical tasks.

**k-Nearest Neighbour**

K-nearest neighbour (KNN) (Fix and Hodges, 1951) is the simplest machine learning algorithm investigated in this thesis. It represents the paradigm of
lazy learning in that no model is built prior to classification of an instance. Instead, when presented with an unlabelled instance lazy algorithms create a generalisation model to classify the instance. For each instance to be classified, the KNN calculates the distance between the given instance and all other instances in the training data. The instance is given the same label as its $K$ nearest neighbours.

**Naive Bayes**

Naive Bayes is a probabilistic classifier which applies Bayes’ Theorem to classify instances. It is naive in the sense that all features in the vector are independent of one another, e.g. for text classification it may treat a text as a bag-of-words, with no account given for the structure or sequences of a given text. While this is a seemingly inherent flaw of the model (Hand and Yu, 2001), it has shown time and again high performance that attributes to its popularity in text classification tasks, such as spam filtering.

**Decision Tree**

Decision Trees are a predictive model where the label of an instance is predicated on the system making decisions on certain feature values, e.g. if the term “no” occurs before the current term, classify the current term as negated. Though there are many implementations of Decision Tree classification algorithms we will investigate the usage of the C4.5 implementation (Quinlan, 1993) which chooses attributes based on the highest normalised information gain.

**Random Forest**

The Random Forest (Breiman, 2001) is a state-of-the-art ensemble classifier. Ensemble learning combines several models to perform classification. At its core, the Random Forest is a collection of decision trees. Its novelty lies in its combination of bagging and feature selection techniques. Bagging (Breiman, 1996) of the training data reduces the issues of over-fitting found in decision trees with feature selection proving robustness to poorly performing features.

**Support Vector Machine**

Another state-of-the-art algorithm, the SVM (Cortes and Vapnik, 1995) represents each instance as an N-dimensional vector, where N is the number of features. The key idea of the SVM is to identify a hyperplane that separates vectors in the data set into the required classes such that the distance between the nearest point from each class to the hyperplane, is a maximum, known as
the maximum margin. Given the high level of dimensionality, the kernel trick i.e. mapping to another feature space, is used to separate instances. The choice of kernel is an important factor and will be further discussed in Chapter 4.

**Sequence Labelling Problems**

Tasks such as POS tagging and Named Entity Detection may be described as sequence labelling problems. While they exhibit many aspects of a classification task, i.e. classifying a token as being an instance of a certain POS class, the classification algorithms described above do not suffice. The key difference being that standalone classifications do not take into account previous classifications, so all structural information is lost. Therefore, algorithms have been developed that leverage previous classifications, such as *Conditional Random Fields* (CRF) and the *Structural Support Vector Machine* (SSVM). Further detail on these algorithms are available in Chapter 3.

**2.2.3 Tools and Resources for Biomedical NLP**

In this section we discuss the tools and resources available to the BioNLP community in order to address domain specific issues which occur such as the huge variation of expression for the same concepts found in clinical texts.

**2.2.3.1 MeSH**

The Medical Subject Headings (MeSH) thesaurus contains a controlled vocabulary of biomedical language, structured in a hierarchy (Rogers, 1963). In this thesaurus the key entity is a descriptor, or subject headings. A descriptor consists of text and a unique ID. The ID defines the descriptor’s placement in the hierarchy, of which it can have several. For example, the concept *Carcinoma, Ductal, Breast* has the unique ID C04.588.180.390. C04 places the concept under *neoplasms*, 588 identifies it as a *localised neoplasm*, 180 highlights the concept under the *breast neoplasms* heading. These kind of relations found in MeSH make it a prime candidate for normalisation of medical vocabulary as well as the identification of synonyms.

**2.2.3.2 UMLS Metathesaurus**

Though a number of thesauri exist in the medical domain, such as MeSH, there is often very little connectivity between them. The Unified Medical
Language System (UMLS) Metathesaurus (Lindberg and Humphreys, 1993) aims to create links between disparate terminology sources available in the biomedical and clinical domains. The Metathesaurus indexes over 100 such collections, creating links and relationship information between entities in the Metathesaurus.

The key entity in the Metathesaurus is the concept. Each concept is identified by a Concept Unique Identifier (CUI), e.g. adrenal gland diseases, adrenal disorder both map to C0001621. The current version of the Metathesaurus contains almost two million concepts. As lexical variants and conceptually linked terms map to the same CUI, it is a valuable asset which allows for the normalisation of terms identified in a text.

2.2.3.3 cTakes

The Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) (Savova et al., 2010) is an open-source natural language processing system to facilitate a range of information extraction techniques on clinical corpora. Its components contain Sentence Boundary detection, tokenisation, normalisation of tokens, POS tagging which extend existing tools to the clinical domain and a Named Entity Recognition system that uses a dictionary look-up function for noun-phrases in a subset of the UMLS Metathesaurus. This also facilitates the semantic normalisation of concepts in that a span may be mapped to a CUI. cTAKES facilitates much of the NLP work required as part of this thesis and as such will also be treated as a baseline in MER and concept normalisation tasks.

2.3 NLP in the Biomedical Domain

Biomedical researchers are expected to keep abreast of current findings, techniques and results in peer-reviewed literature. However, given that almost 1,800 new papers are published daily (Hunter and Cohen, 2006), this task is unfeasible to be undertaken by a single individual. Therefore NLP and IR techniques are required to retrieve, extract and present the required information in an easy-to-digest manner. In this section, we present an overview of NLP techniques developed on biomedical literature as well as datasets and shared tasks that accelerated research in the area.
A pipeline for Biomedical NLP

Before any high-level form of NLP, (such as summarisation or information extraction) may take place, texts must go through a set of more rudimentary processes. These processes include tokenisation, POS-tagging, dependency analysis and semantic analysis.

The first stage of the process requires sentences to be delimited through the use of a *sentence splitter*. Detecting a sentence boundary is not a trivial issue as punctuation marks such as a period often do not mark the end of a sentence, therefore standard pattern-matching systems do not suffice. Instead systems such as GENIASS\(^1\) employ a machine-learning based approach, training a maximum entropy classifier to classify a set of candidate delimiters using features such as n-grams before and after the token. GENIASS has been trained on 2000 MedLine abstracts from the GENIA corpus (Kim et al., 2003) achieving an F-score of 0.997.

*Tokenisation* is then performed on the split corpus. Rule-based systems have been found to generalise poorly (He and Kayaalp, 2006) so machine-learning approaches are often implemented for use in the biomedical domain achieving f-scores greater than 0.90 (Barrett and Weber-Jahnke, 2010).

*POS-tagging* allows for the incorporation of syntactic information into language processing workflows. Machine-learning approaches again surpass the use of rule-based systems. Systems have been evaluated on GENIA and PennBIOIE corpora (Tsuruoka et al., 2005) with the state-of-the-art approaches achieving f-scores greater than 0.97 (Smith et al., 2003). Tsuruoka et al. (2005) have shown that machine-learning approaches to POS-tagging generalise well if the source data domain matches that of the target domain.

Sentence splitting, tokenisation and POS-tagging are pre-processing problems often encountered in the general domain. Named Entity Recognition (NER) in the biomedical domain is a non-trivial problem that forms the basis of any information extraction or reasoning system. Unlike the general domain where the task usually requires the discovery of *locations, individuals* or *companies*, biomedical named entity recognition is concerned with the discovery of biomedical entities such as genes, proteins, cells and drugs. This task is more difficult to generalise entities than in the general domain by the nature of named entities in biomedical literature. General domain entities usually abide by certain conventions such as capitalisation. Biomedical literature is rich in synonymy

\(^{1}\)http://www.nactem.ac.uk/y-matsu/geniass/
and spelling variation, missing conventions found in the general domain. For example, as there is no set criteria in defining a gene name e.g. *caspase3* and *CASP3* or *NF-kappaB* and *NFKappaB*. We will now discuss tasks specific to the biomedical domain, and how their research is promoted through shared tasks and datasets in the community.

**Shared Tasks in Biomedical NLP**

The first biomedical NLP challenge to incorporate NER took place in 2002 as part of the KDD Challenge Cup\(^2\). Task 1 of the challenge required the extraction of genes mentioned in a subset of MEDLINE papers. Following the success of this challenge, BioCreAtIvE (Critical Assessment of Information Extraction in Biology)\(^3\) set out to evaluate information extraction in the biology and biomedical domain. The first task of the challenge focused on gene name extraction their normalisation. Again, MEDLINE abstracts and full text papers were used to evaluate systems. Top performing systems achieved f-scores 0.84 using machine learning approaches, outperforming rule-based submissions.

BioNLP began as a challenge in the recognition of five biomedical named entities (*protein, DNA, RNA, cell line, cell type*)\(^4\). Evaluations were performed using the GENIA corpus and 400 newly annotated abstracts. Unfortunately, systems submitted to the initial challenge achieved significantly lower performance, a fact attributed to inconsistencies in gold standard annotation (Dingare et al., 2005).

Following the high level of interest among researchers in biomedical text mining both, BioNLP and BioCreAtIvE shared tasks have continued to the present day. BioCreAtIvE has continued its focus on biomedical entity recognition, though entities have been broadened to drug and chemical compound recognition (BioCreAtIvE IV) and the relations between entities, such as protein-protein interactions (BioCreAtIvE II). Similarly, BioNLP has extended its challenges to event extraction, negation and speculation detection (BioNLP 2009), coreference and relation detection (BioNLP 2011) as well as further event types (BioNLP 2013).

These tasks aid the development in two ways. First, through the tasks themselves. However, following the completion of tasks data is often made freely available allowing for further development of systems. BANNER (Leaman and Gonzalez, 2008), ABNER (Settles, 2004) and other top performing systems

\(^2\)http://www.kdd.org/kddcup/index.php
\(^3\)http://www.biocreative.org/
\(^4\)http://www.nactem.ac.uk/tsujii/GENIA/ERtask/report.html
often implement machine learning based methods such as CRFs, mirroring approaches from the general domain where rule-based systems do not attain satisfactory performance.

In this section, we presented an overview of challenges for NLP in the biomedical domain. These challenges and shared tasks facilitate the development of systems that address domain-specific issues found in biomedical texts. More and more systems adopt machine learning techniques as the solution to NLP tasks, ranging from sentence splitting and tokenisation to more complex NER tasks, achieving results similar to those in the general domain. The development and research of new techniques has been aided greatly by datasets being made available through challenges such as BioCreAtIvE.

Despite interesting results, Collier et al. (2005) note that clinical corpora, such as electronic health records, should be used to bridge the gap between bioinformatics and medical informatics. In the next section, we provide an overview of NLP problems identified in the clinical domain and methods to resolve the issues outlined in this section.

### 2.4 NLP in the Clinical Domain

Documentation in healthcare is becoming increasingly digitised, as shown by the prevalence of Electronic Health Records (EHR) described in Chapter 1. Much like the biomedical literature discussed in Section 2.3, clinical data is growing at an explosive rate with multiple reports generated per patient visit. For example, on the admission of a patient, an emergency room report may be filed, followed by History & Physical, Surgical Pathology or Radiology Reports, on leaving an institution a discharge summary is created. As a result, research in clinical NLP has focused on methods to extract and reason about pertinent data contained in clinical reports with the overall goal of aiding the healthcare process, e.g. clinical decision support systems (Demner-Fushman et al., 2009).

**Data-Sharing issues in the Clinical Domain**

The domain of clinical NLP has garnered a large amount of interest in the research community given its promise in advancing the world of healthcare (Wu et al., 2013). In spite of this, the sensitive nature of clinical data causes difficulties in distributing datasets thus hindering research in the domain. *It is key that a patient’s identity cannot be discovered, therefore several security and ethical*
measures must be enforced to safeguard the patient’s privacy. In the past, methods of de-identification either were not available, or not good enough for healthcare institutions to willingly distribute data. More recently de-identification systems as well as security and legal measures have been developed to ensure privacy of individuals. As a result, data can now be distributed subject to license agreements and ethics approval. The development of these strategies such as the BLULab at the University of Pittsburgh allowed individual research teams to evaluate their systems on clinical corpora.

An NLP pipeline in the clinical domain

A common trait throughout the years of research in clinical NLP is the investigation of adding structure to unstructured clinical reports (Hirschman et al., 1976) with varying applications, from creating a structured format for retrieval (Sager et al., 1977) to natural language understanding and reasoning systems (Christensen, 2002).

To facilitate research at the top end of the NLP pipeline, again low-level processing such as POS-tagging & sentence-splitting are required. Similar to other domains of NLP, machine learning approaches are adopted. Commonly, approaches to sentence splitting achieve satisfactory performance using machine-learned models trained on MEDLINE data, such as GENIASS. However, as Pakhomov et al. (2006) show tasks such as POS-tagging require data which is of the same domain. Testing of POS-tagging across report types has shown performance below that found in the general domain, achieving f-scores of 0.88 when evaluated on unseen data (Fan et al., 2011) suggesting that annotations rather than models should be shared. However, this again faces issues of data distribution in the clinical domain. To work around this issue Ferraro et al. (2013) investigate using a domain adaptation technique (Daume III, 2009) that combines annotated data from the general domain with text from the clinical domain. The technique showed great improvements, reporting f-scores up to 0.939.

Initial methods in Medical Entity Recognition

MetaMap (Aronson and Lang, 2010) laid the foundations for the burgeoning research area of NER in the clinical domain. MetaMap is a rule-based algorithm designed to mine the UMLS Metathesaurus in order to map textual spans to concept identifiers in the Metathesaurus. However, since its original conception in 2001 (Aronson, 2001), its runtimes and memory requirements have increased drastically owing to the constant expansion of the UMLS Meta-
Thesaurus. Despite respectable performance in tasks such as query expansion for Information Retrieval, its disadvantages have led researchers to develop more efficient methods.

The IndexFinder algorithm (Zou et al., 2003) increases the efficiency of mapping strings to UMLS concepts through treating the UMLS as an inverted index. However, this approach does not identify entity spans, rather signifies the given text may map to a set of CUIS. Ruch et al. (2003) posits a system that combines spell-checking, POS tags and word sense disambiguation techniques to recognise patient and doctor names. Kipper-Schuler et al. (2008) uses a dictionary lookup system that, while efficient, achieves an f-score of only 0.56. Low performance scores were common among rule-based and dictionary lookup systems leading the community to explore options in machine learning. Li et al. (2008) moved towards disease & disorder recognition through Conditional Random Fields and Support Vector Machines, with a CRF based approach achieving an f-score of 0.86. Similarly, Wang and Patrick (2009) combined SVM and CRF approaches through a cascade system achieving an f-score of 0.833 across 10 entity types. Though interest in this domain was on the rise, the unavailability of clinical datasets stifled research in that results across research groups could not be compared. This also reduced the ability to focus and define the NER task for the clinical domain.

**Reasoning of deeper knowledge of Medical Conditions**

Though the identification of entities is useful, research in the community has extended to developing systems that perform deeper analysis of recognised entities. This analysis takes the form of negation detection, as well as assertion classification. As described briefly in Chapter 2, the assertion status of a medical condition denotes how it may affect the patient or related individual. For example, the sequence *No eczema* describes the condition eczema with a negated assertion status. The problem of assertion classification is discussed in more detail in Chapter 4.

The first and perhaps most notable foray into negation detection was the Negex algorithm (Chapman et al., 2001). This rule-based algorithm checks for the occurrence of certain *trigger* terms within the span of the inspected concept, achieving a specificity of 94.5%. This algorithm was modified by Chapman et al. (2007) in order to recognise further assertion statuses. It determined negation, temporality (recent, historical, hypothetical) and experiencer (patient, other). This approach worked well in negation detection (f-score 0.97) and hypothetical conditions (f-score 0.88). However, it did not fair as well in de-
termining historical conditions (0.71) or experiencer information (f-score 0.67). Further evaluations in Harkema et al. (2009) across report types reaffirmed previous experimental findings. As a result of low scores on further assertion statuses, research in the area began looking at more machine learning based methods that make use of information other than lexical items. Comparative evaluations of machine learning and statistical systems with respect to rule based systems have seen machine-learning systems significantly outperform rule-base approaches (Mowery et al., 2009; Uzuner et al., 2009). However, results among systems are not comparable between research groups due to the inability to distribute annotations.

Shared tasks in evaluating NLP research have provided a means to ethically distribute clinical data in order to comprehensively and comparatively evaluate research in the area. The i2b2 (Informatics for Integrating Biology and the Bedside) group accelerated research in clinical NLP through its organisation of shared tasks among the community. These tasks are presented in Table 2.1. The goal of this thesis is to investigate the applicability of NLP tasks in resolving issues of term dependencies, assertion status in querying and vocabulary mismatch. With respect to the work described in this thesis, the 2010 Challenge is most relevant in its investigation of Named Entity Recognition and Assertion Classification. We therefore focus on these tasks as part of the challenge.

<table>
<thead>
<tr>
<th>Year</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>Obesity Challenge</td>
</tr>
<tr>
<td>2009</td>
<td>Medication Extraction</td>
</tr>
<tr>
<td>2010</td>
<td>- Named Entity Extraction</td>
</tr>
<tr>
<td></td>
<td>- Assertion Classification</td>
</tr>
<tr>
<td></td>
<td>- Relation Detection</td>
</tr>
<tr>
<td>2011</td>
<td>Co-reference resolution</td>
</tr>
</tbody>
</table>

Table 2.1: List of i2b2 Shared Tasks 2008-2011

In the next sections, we will focus on two tasks whose datasets were used to evaluate the systems produced in this work. These tasks were part of the i2b2/VA Challenge 2010 as shown in Table 2.1 and the 2013 ShARe/CLEF eHealth Evaluation Lab.

## 2.5 Approaches at the Fourth i2b2 Challenge

The fourth i2b2/VA challenge presented three tasks in the extraction of medical concepts and supplementary information.
These three tasks were as follows:

1. **Concept Identification**: The extraction of medical concepts found in free-text. These concepts may be a medical problem, a test or treatment.

2. **Assertion Classification**: The classification of the *Assertion Status* of a medical problem.

3. **Relation Detection**: The identification and classification of relations that may hold between identified concepts.

For these tasks Partners Healthcare, along with Beth Israel Deaconess Medical Centre and the University of Pittsburgh Medical Centre released de-identified discharge summaries on the condition of a Data Use Agreement through the i2b2 challenge. In November 2011 this data was released to the wider research community subject to Data Use Agreements. All annotations of concepts, assertions and relations in these datasets were performed by healthcare professionals.

For the purpose of this work, we focus on the tasks of Concept Identification and Assertion Classification. 22 systems were submitted for the Concept Identification task, with 21 assertion classification systems submitted for evaluation under the challenge. We will now discuss the submitted systems.

### 2.5.1 Concept Identification

Overall 22 systems were submitted for the Concept Identification task. The top ten performing systems all used machine-learning approaches through supervised, semi-supervised or hybrid methods, as shown in Table 2.2. Supervised methods, in particular Conditional Random Fields (CRF), were by far the most widely used approach. The high usage of CRFs reflects that of sequence labelling problems across domains, as it avoids the *label bias* problem found in other Markov approaches. The label bias problem states that observations with low entropy are ignored given a higher entropy sequence of states. Given that CRFs address this issue, it is often a common approach in state-of-the-art systems. Despite their suitability and high performance, CRFs require a considerable amount of time to train. This drain on processing time is a cause of concern for researchers, and incentivises the exploration of other supervised methods. In the task, CRFs were combined with different training
methods such as Maximum Entropy methods (MEMM) (Torii and Liu, 2010) as well as other classifiers such as the high performance Support Vector Machine (SVM) (Jiang et al., 2010; Patrick et al., 2010; Roberts et al., 2010). The state-of-the-art system submitted (de Bruijn et al., 2010) realised the potential of CRFs while minimising training effort through the use of semi-Markov methods. Semi-Markov models label multi-token spans rather than single tokens allowing for a richer feature-set, e.g. the number of tokens in the span. The process of training the models was performed using the PA algorithm (Crammer et al., 2006) thus further shortening the training time.

Though these systems achieve high-performance, they are often highly complex and require significant preprocessing. A third approach removes some of these complexities through combining existing systems and libraries (Kang et al., 2010). Though seemingly simplistic, the third performing system in the challenge uses a voting based approach among six systems to achieve higher performance than any single system.

Feature sets used in recognising medical entities closely match those used in the general domain, shown in Table 2.3. The use of token-based features is common. These features extract information about a given token, such as prefix/suffix information as well as orthographic information such as capitalisation and the presence of hyphens. Throughout the systems, windowing functions and POS features are employed. POS features indicate the part-of-speech for a given token. A windowing function extracts features for tokens that surround the token currently being inspected, giving more contextual information about the given token.

Semantic information relating to a token also was an important feature for the task. To extract semantic information two approaches were taken. The first was to use a tool such as cTakes, MetaMap or ProMiner to generate normalised concepts. The second approach was to use a gazetteer list, that although effective did not match the approaches using specialised tools.

The task also introduced more novel feature-sets. Jonnalagadda and Gonzalez (2010) introduced the use of similarity vectors as features to determine if a token’s distribution may impact its likelihood of being a concept. Similarly de Bruijn et al. (2010) generated clusters using the Brown clustering algorithm (Brown et al., 1992). The path to the hierarchical cluster to which a token was a member was then used as a feature. Further contextual information such the section header under which a token occurs may also be extracted as a feature (Torii and Liu, 2010).
<table>
<thead>
<tr>
<th>System By</th>
<th>Semi-Markov</th>
<th>CRF</th>
<th>SVM</th>
<th>Ensemble</th>
<th>MEMM</th>
<th>Exact Match</th>
<th>Inexact Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>de Bruijn et al. (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.852</td>
<td>0.924</td>
</tr>
<tr>
<td>Jiang et al. (2010)</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>0.839</td>
<td>0.913</td>
</tr>
<tr>
<td>Kang et al. (2010)</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>0.821</td>
<td>0.904</td>
</tr>
<tr>
<td>Gurulingappa et al. (2010)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>0.818</td>
<td>0.905</td>
</tr>
<tr>
<td>Patrick et al. (2010)</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>0.818</td>
<td>0.898</td>
</tr>
<tr>
<td>Torii and Liu (2010)</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>0.813</td>
<td>0.898</td>
</tr>
<tr>
<td>Jonnalagadda and Gonzalez (2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.809</td>
<td>0.901</td>
</tr>
<tr>
<td>Sasaki et al. (2010)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>0.802</td>
<td>0.887</td>
</tr>
<tr>
<td>Roberts et al. (2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.796</td>
<td>0.893</td>
</tr>
<tr>
<td>Pai et al. (2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.788</td>
<td>0.884</td>
</tr>
</tbody>
</table>

Table 2.2: Top Concept Identification Systems at i2b2 2010: Machine Learning Algorithms
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>de Bruijn et al. (2010)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Jiang et al. (2010)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kang et al. (2010)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gurulingappa et al. (2010)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patrick et al. (2010)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Torii and Liu (2010)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jonnalagadda and Gonzalez (2010)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sasaki et al. (2010)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roberts et al. (2010)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pai et al. (2010)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: Top Concept Identification Systems at i2b2 2010: Feature-Sets
Beyond the I2B2/VA Challenge

Following the completion of the task, several further works have been published building on systems presented in the i2b2/VA task. Abacha and Zweigenbaum (2011) comparatively investigates the use of semantic and statistical approaches to the issue. It featured an in-depth evaluation using MetaMap, chunking based methods and statistical approaches leveraging conditional random fields. Their experiments found that a hybrid system combining semantic and statistical approaches faired best on the i2b2/VA corpus, though its performance did not match the state-of-the-art at 2010 i2b2/VA challenge.

The most recent work on Medical Entity Recognition on the i2b2/VA corpus presented in (Tang et al., 2012) outperforms the state-of-the-art at the i2b2 challenge. Its implementation bears many similarities to de Bruijn et al. (2010) with two important differences. First, a change in classification algorithm. While many tagging systems at the i2b2/VA focused on the usage of CRFs or other Maximum Likelihood Estimation (MLE) approaches the authors investigate the usage of Structural Support Vector Machines (SSVM) that model the problem by the large margin method allowing greater generalisability. The second change to most standard approaches is the use of the BIESO (Beginning, Intermediate, End, Single-Word, Outside) tagging format rather than the more traditional BIO format. The overlying goal of this approach is to allow for more differentiation between single-word and multi-word expressions. The BIESO approach is described in more detail, with experimental analysis in Chapter 3. Experimental results in the literature show that SSVMs require less training time than CRFs as well as outperform CRFs in a series of tasks, the introduction of the BIESO format also saw gains in performance to achieve an f-score 0.05 higher than leading systems at the i2b2/VA challenge.

2.5.2 Assertion Classification

There were 21 systems submitted for the task of Assertion classification. Similar to the trend in the Concept Identification task, the ten top performing systems integrated machine learning in some form, most often through standalone machine learning approaches, though hybrid systems are also investigated. The approaches used are given in Table 2.4, with an overview of their respective performance. The top-performing system for this task was developed by de Bruijn et al. (2010), again owing to the large number of features and novel approach to training systems. The SVM is the most commonly employed ap-
approach in classification tasks as it often provides high performance when optimally trained.

CRF-based approaches also scored highly (Clark et al., 2010; Jiang et al., 2010) using Maximum Entropy methods as well as in combination with SVM approaches. Chang et al. (2010) combined an array of approaches to achieve high performance rather than relying on one machine-learning technique.

<table>
<thead>
<tr>
<th>System By</th>
<th>Semi-Markov</th>
<th>CRF</th>
<th>SVM</th>
<th>ME</th>
<th>Log. Reg.</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>de Bruijn et al. (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.936</td>
</tr>
<tr>
<td>Clark et al. (2010)</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>0.934</td>
</tr>
<tr>
<td>Demner-Fushman et al. (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.933</td>
</tr>
<tr>
<td>Jiang et al. (2010)</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>0.931</td>
</tr>
<tr>
<td>Grouin et al. (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.931</td>
</tr>
<tr>
<td>Divita et al. (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.930</td>
</tr>
<tr>
<td>Cohen et al. (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.928</td>
</tr>
<tr>
<td>Roberts et al. (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.928</td>
</tr>
<tr>
<td>Anick et al. (2010)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.923</td>
</tr>
<tr>
<td>Chang et al. (2010)</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>0.921</td>
</tr>
</tbody>
</table>

Table 2.4: Top Assertion Classification Systems at i2b2 2010: Machine Learning Algorithms

Although machine learning systems were the most common approach, they often implemented rule-based features, i.e. machine learning features that are generated from patterns in text. An overview of the implemented feature-sets is given in Table 2.5. The output of the rule-based algorithm ConText (Chapman et al., 2007) is used in 3 of the top 5 systems. The occurrence of trigger or cue words provide insight to the surroundings of a concept.

The top 5 systems all use section header information that can greatly improve performance e.g. concepts occurring under the header family history often relate to someone other than the patient. Similar to the Concept Identification task, information about the token, its part-of-speech and information relating to surrounding tokens are a common occurrence among feature-sets. The top performing systems also extract normalised semantic information, using tools such as MetaMap and cTakes. Two systems use deeper syntactic information (Cohen et al., 2010; Chang et al., 2010) rather than relying solely on POS information. Cohen et al. (2010) also use metric information such as term-frequency scores to determine the correct assertion.
<table>
<thead>
<tr>
<th>System By</th>
<th>Token</th>
<th>Window</th>
<th>Section</th>
<th>Doc.</th>
<th>POS</th>
<th>Sem.</th>
<th>Syn.</th>
<th>Trig.</th>
<th>ConText</th>
<th>Rule</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>de Bruijn et al. (2010)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clark et al. (2010)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demner-Fushman et al. (2010)</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jiang et al. (2010)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grouin et al. (2010)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Divita et al. (2010)</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohen et al. (2010)</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roberts et al. (2010)</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anick et al. (2010)</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chang et al. (2010)</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.5: Top Assertion Classification Systems at i2b2 2010: Feature-Sets
Beyond the I2B2/VA Assertion Classification Task

Following the i2b2/VA Assertion Classification task, systems have been refined to focus on feature-engineering to achieve state-of-the-art performance. Kim et al. (2011b) extend their previous approach through adapting the multi-class learning strategy, redundancy removal from training data and new features that focus on increasing the recall on minority classes. The authors used a One-Versus-All (OVA) approach rather than a One-Versus-One (OVO) strategy on the supposition that doesn’t work well for highly skewed datasets (Huang et al., 2006). The adjustments to the system significantly improved performance to that of state of the art, though minority classes still experience relatively poor performance.

Bejan et al. (2013) further investigate the resolution of assertion status in the context of phenotype identification. Their system introduces features that analyse if a given concept is under the focus of an assertion keyword as well as features that identify keywords within a certain span most associated with assertion types. Classification was performed using a multi-class SVM. The output of this approach significantly outperformed the system of Kim et al. (2011b).

2.6 CLEF/ShARe eHealth 2013

The 2013 CLEF/ShARe eHealth Evaluation Lab (Suominen et al., 2013) focused on natural language processing (NLP) and information retrieval (IR) for clinical care. The Evaluation Lab featured three tasks as listed below:

1. **Named Entity Recognition and Normalisation of disorders**: Recognition of disorders in clinical records and mapping to a UMLS concept identifier (CUI).

2. **Normalisation of acronyms/abbreviations**: Recognition and expansion of abbreviations and acronyms in clinical text, they are then to be mapped to a UMLS CUI.

3. **Information retrieval to address questions patients may have when reading clinical reports**: TREC style information retrieval tasks. This differs to the TREC Medical track in that web-based documents, rather than clinical reports are retrieved in order to answer patient questions.
For this work we focus on Task 1 as part of the evaluation challenge. Task 1 has two aims. First, the recognition of medical entity spans within clinical reports that belong to the Disorders semantic group in the UMLS. The second aim is to map a recognised entity to a concept identifier in the UMLS, thus normalising the concept. Given the high level of synonymy and lexical variation found in clinical texts, normalisation has great potential in aiding further tasks in an NLP or IR pipeline.

The recognition task differs to the Fourth i2b2/VA challenge in that recognised entities may be disjoint. A disjoint entity span may be broken or disjointed by non-entity terms. For example, in the sentence below the tokens left, atrium and dilated map to the concept identifier C0344720 in the UMLS. Rather than incorrectly map these terms independently to separate, and more than likely incorrect CUIs, they are recognised as one entity separated by the token span is moderately.

- The left atrium is moderately dilated.

At the challenge a range of methodologies were developed for this novel task. In the detection of medical entities, approaches frequently employed machine learning based approaches such as CRFs (Patrick et al., 2013), SSVMs (Tang et al., 2013) and freely available systems such as cTakes (Liu et al., 2013) and BANNER (Leaman et al., 2013).

In the case of concept normalisation at CLEF 2013, top performing systems relied heavily on MetaMap with a combination of rules (Osborne et al., 2013; Zuccon et al., 2013) while other approaches include rule-based lookup (Liu et al., 2013) and Vector-Space indexing of the UMLS (Tang et al., 2013). Chapter 3 will feature further detail on the author’s system at CLEF 2013.

### 2.7 Task Definitions

In this section, we will define the tasks which will be investigated as part of this thesis.

**Concept Identification at the Fourth i2b2/VA Challenge**

For the Concept Identification task at the Fourth i2b2/VA challenge, there are three entity types to be recognised: problem, treatment and test. Systems must identify a concept span and correctly assign the span to one of the three possible entity types. Evaluation of the task allows exact and inexact matching.
Exact matching requires the identified span matches that of the gold standard, where inexact matching requires that spans need only overlap.

**Disorder Recognition and Concept Normalisation at CLEF/ShARe 2013**

Disorder Recognition at CLEF/ShARe follows the same template as the i2b2/VA Challenge in that an entity must be identified in a clinical document. The task differs as there is only one entity type to be recognised, namely: disorder. For all intents and purposes, disorder concepts are equivalent to a problem concept. The task deviates further through the normalisation of concepts. Concept normalisation is performed through mapping of an identified concept to a CUI in the UMLS Metathesaurus.

**Condition Attribution and Temporal Grounding**

The Assertion Status of a medical condition denotes the extent to which the condition affects the patient, if at all. The task of Condition Attribution determines if a given medical condition applies to the patient directly, or to another individual such as a family member. Temporal Grounding resolves whether a medical condition is currently impacting a patient, or is described as part of the patient’s history.

**Assertion Classification at the Fourth i2b2/VA Challenge**

Assertion Classification at the Fourth i2b2/VA challenge requires six possible assertion statuses. The assertion statuses are as follows: present, absent, hypothetical, possible, conditional and not associated with the patient (NotPatient). Instances of NotPatient are equivalent to negative instances in the Condition Attribution task. Though no assertion status in the i2b2/VA challenge directly relates to the Temporal Grounding task, the assertion status Absent is defined as negation but also historical instances that no longer affect the patient.

With more possible classes, this becomes a multi-class classification problem, rather than a binary system in the case of Condition Attribution and Temporal Grounding.

### 2.8 Summary

In this chapter, we provided an overview of the theoretical background of Natural Language Processing, as well as a discussion of methodologies and evaluations featured in the literature. In relation to our aims for this chapter, we have discovered the answers to the following research questions:
What machine learning approaches perform best in Medical Entity Recognition?

Throughout the literature, and especially the i2b2/VA challenge we see the CRF as the preferred technique for the task. However, given high computational requirements of CRFs, researchers in the domain have begun to explore other options such as Semi-Markov models and Structural Support Vector Machines. While Semi-Markov approaches performed best at the i2b2/VA challenge, it has been bested by the Structural Support Vector Machines in post-challenge experiments, leading the SSVM to become the state-of-the-art approach to Medical Entity Recognition. For this reason, as well as its improved training times over CRF based techniques, the work in this thesis leverages the Structural Support Vector Machine.

What features best aid the task of Medical Entity Recognition?

Features in the task of Medical Entity Recognition largely follow that of the general domain, i.e the use of lexical, POS and token information within a window. However, some domain specific features are introduced which sees gains in performance. For example, semantic information from the UMLS Metathesaurus was used in three of the top four systems at the i2b2/VA challenge. Furthermore, state of the art approaches also consistently leverage hierarchical clustering information (de Bruijn et al., 2010; Tang et al., 2012).

What machine learning approaches perform best in Assertion Classification?

Approaches at the i2b2/VA challenge overwhelmingly chose the Support Vector Machine as a classification algorithm. Although the Semi-Markov approach achieved top performance, the difference in system performance is marginal. Due to its relative ease of understanding, efficient implementation and consistently high performance in classification tasks the SVM will be further investigated in this thesis for the multi-class task. However, further experimentation with other classifiers will also be examined.

What features best aid the task of Assertion Classification?

While NER in the general domain provides a set of features that form the basis of Medical Entity Recognition, there are no general domain tasks which directly relate to the problem of Assertion Classification. Given the task’s novelty, there is much more variance in the features used in the classification task. A mainstay among i2b2/VA is the use of token information within a window, with all but two of the top ten performing approaches using this information.
Notably, the use of semantic and header information is seen across all the top performing approaches.

To summarise, in the literature we have seen an array of machine-learning techniques and feature-sets designed to address the issues of Medical Entity Recognition and Assertion Classification. However, the literature is not conclusive in addressing what contributes most to each task. However, we may note the contribution of Structural SVM in the concept identification task, as well as the use of semantic and clustering information. The SVM is also commonly implemented in the Assertion Classification task, where we see token-based information as well as rule-based patterns contribute as features. Throughout the rest of this thesis, experiments are performed to determine algorithms and feature sets best suited to the task based on our findings in the literature.

In the next two chapters we will build on our findings as a result of the discussion in this chapter. In Chapter 3, we evaluate a Medical Entity Recognition system that implements many of the features discussed. In particular, we focus on determining which features contribute most to the task. In Chapter 4 we turn our attention to the task of Assertion Classification where we also aim to discover the most contributory feature sets and algorithms.
CHAPTER THREE

NAMED ENTITY RECOGNITION AND CONCEPT NORMALISATION IN THE MEDICAL DOMAIN: EXPERIMENTS

3.1 Introduction

In Chapter 2, we discussed tools and methods available to facilitate the creation of solutions to natural language processing tasks. In this chapter, we apply these technologies to the task of named entity recognition in the medical domain and concept normalisation. Though in the medical domain it remains an entity recognition task much like in the general domain, the term Medical Entity Recognition (MER) is a more accurate portrayal of the task as its focus lies on medical entities such as problems and treatments, rather than named entities such as locations or individuals. Thus, throughout this thesis the task will be described as Medical Entity Recognition. Concept Normalization is the task of mapping a span of text to a unique identifier in a domain-specific ontology. The use of concept normalisation has many application areas in NLP and IR as it reduces the negative effects of rich synonymy and linguistic variation.

In this chapter, we present two machine learning based MER systems, which are evaluated on datasets from the Fourth i2b2/VA Challenge and the CLEF/ShARe evaluation lab. We also discuss a concept normalisation system that leverages the MER system’s output to recognise concept spans prior to mapping them to an ontology. The system is evaluated using the CLEF/ShARe challenge corpus and gold standard annotations. These systems and their development are used to investigate the following research questions:
• Given top-performing feature sets in the literature, we ask what features contribute the most to the task, with the aim of removing redundant information and improving performance.

• What tagging format best suits the recognition of continuous and disjoint entities?

• What are the issues in extending standard MER approaches to the recognition of disjoint entities? What features are best suited to this task?

• To develop an MER system that is comparable to the current literature both in its implementation and evaluation.

• What are the key issues in concept normalisation?

In Section 3.2 we set out the metrics used to evaluate the systems. Section 3.3 provides a description of an initial MER system, trained and evaluated on the i2b2/VA challenge dataset. In Section 3.4, we adapt the system from Section 3.3 so that it may be evaluated on the CLEF/ShARe dataset. This set of experiments has several goals. First, to develop a medical entity recognition system comparable to that found in current literature. In doing so, further approaches may build on this work, both on these challenge datasets and their application to the information retrieval domain. Furthermore, we analyse the contributions of feature sets through the combination of novel and more widely used feature sets. Following our feature analysis of the MER systems, we will discuss and evaluate approaches to Concept Normalisation in Section 3.4.3.2 on the CLEF/ShARe dataset, that leverage the output of the MER system from Section 3.4. This is summarised in Figure 3.1. The overall goal of these systems are the following: MER provides the ability to automatically detect phrasal boundaries in natural language queries and texts, as useful trait where clinical concepts are often multi-word items. Secondly, Concept Normalisation provides a manner to address vocabulary mismatch through mapping concepts in documents and queries to a domain-specific ontology.

### 3.2 Evaluation Metrics

In this section, we provide an overview of the metrics used to evaluate the MER and Concept Normalization systems described in this chapter. To evaluate the techniques we use the widely implemented metrics of Precision, Recall
and $F$-score. These IR-based measures, used in the shared task community, must be defined in terms of a classification task. First we define true positives, false positives and false negatives as in Table 3.1.

<table>
<thead>
<tr>
<th>System Prediction</th>
<th>Human Annotation</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>True Positive (TP)</td>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>Negative</td>
<td>False Negative (FN)</td>
<td>Negative</td>
</tr>
</tbody>
</table>

Table 3.1: Calculating TP, FP and FN for Machine Learning Evaluation

From these statistics, Precision and Recall can be calculated using these measurements in Equations 3.1 and 3.2, respectively. $F$-score is calculated as the harmonic mean of precision and recall in Equation 3.3. All three measurements lie between 1 and 0, where 1 indicates best performance.

$$Precision(P) = \frac{TP}{(TP + FP)} \quad (3.1)$$

$$Recall(R) = \frac{TP}{(TP + FN)} \quad (3.2)$$

$$F_{score} = \frac{2 \times P \times R}{(P + R)} \quad (3.3)$$

For the Concept Identification task we allow for exact and inexact measurements as found in the i2b2/VA 2010 challenge (Uzuner et al., 2011) and the
CLEF/ShARe tasks. Exact measurements require that the bounds of the identified span and the concept type, e.g. *Problem, Test*, match exactly with that of the gold standard. An inexact metric relaxes the exact span requirement so that the gold standard and predicted span need only overlap and have matching concept types to be counted as a true positive.

As the Concept Identification contains three classes (problem, treatment and test) and the Assertion Classification task has six (present, absent, possible, conditional, hypothetical, not associated with the patient) we must discuss how best to evaluate the performance of systems across all classifications. Often, average f-scores across all classifications are computed. There are two methodologies to compute these average scores. The first is known as *macro-averaging*. This method first calculates the average of the precision and recall of the system for each classification task. The new scores are then combined as in Equation 3.3 to create the macro-average f-score. *Micro-averaging* produces an average f-score through summing true positives, false positives and false negatives. These new rates are used in the equations outlined above to produce the micro-average f-score. Macro-averaging gives equal weighting across all possible classes as an indicator of overall performance. This equal weighting does not suffice where there is a high data skew. Typically, where a class has very few positive instances, classification performance is often poor. With equal weighting, this poor performance will severely impact overall performance. As such, micro-averaging gives a clearer picture of performance by taking into account class frequency. Throughout our evaluations we will report on both micro and macro averaging where class skew is an issue.

The final metric is applied to the Concept Normalisation. This metric of Accuracy was used in the CLEF/ShARe Evaluation Lab for the task of Concept Normalisation. It is defined in Equation 3.4.

\[
\text{Accuracy} = \frac{c}{t} 
\]  

(3.4)

where \(c\) is the number of predicted disorders with correct span and CUI, \(t\) = Total number of annotated disorders in the gold standard

As with other metrics defined, there is both an exact and inexact definition of Accuracy. For the inexact version, the span requirement is relaxed such that only an overlap between the prediction and gold standard is required.
3.3 The Fourth i2b2/VA Challenge

In this section, we provide a discussion of an MER system developed using the available subset of the Fourth i2b2/VA Challenge dataset. The challenge requires participants to develop systems that can identify the spans of medical problems, tests and treatments in clinical text. In Section 3.3.2 we present an overview of the development of the system. Section 3.3.3 evaluates the performance of the MER approach on the i2b2/VA corpus, in particular we focus on feature contributions.

3.3.1 Dataset Overview

The 2010 i2b2/VA Challenge continued i2b2’s (Uzuner et al., 2011) efforts in fostering NLP research in the clinical domain through releasing clinical reports in an ethical manner through shared tasks as discussed in Chapter 2. For the 2010 task Partners Healthcare, along with Beth Israel Deaconess Medical Centre and the University of Pittsburgh Medical Centre released de-identified discharge summaries on the condition of a Data Use Agreement. In November 2011 this data was released to the wider research community subject to Data Use Agreements. The dataset comprised of 172 annotated training reports, 256 test reports as well as a further 268 unannotated reports.

The 2010 task had three tiers of tasks: Concept Identification, Assertion Classification and Relation Detection. This thesis focuses on Concept Identification and Assertion Classification. The Concept Identification task requires the identification of spans that contain concepts. These concepts may fit into the subgroups of problem, treatment and test. Examples of these concepts are provided in Table 3.2.

<table>
<thead>
<tr>
<th>Problem</th>
<th>Treatment</th>
<th>Test</th>
</tr>
</thead>
</table>
| There are simple atheroma in the aortic arch. | I suggest broad spectrum antibiotics. | Echo was obtained which showed EF of 35-40%.

Table 3.2: Examples of Problem, Treatment and Test concepts in the i2b2 Challenge Corpus

A breakdown of the annotated concepts is presented in Table 3.3. Concepts of type Problem are by far the most numerous accounting for over 7,000 and 12,000 instances in the training and test set, respectively. The number of test instances is between 1.5 and 2.5 times the size of the training instances. This
increase in size replicates the increase in data size when a system is run in a real world application, allowing for some generalisations in analysing its performance. In the next section we describe the system developed and evaluated using this dataset.

<table>
<thead>
<tr>
<th></th>
<th>Problem</th>
<th>Treatment</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>7,073</td>
<td>4,844</td>
<td>4,608</td>
</tr>
<tr>
<td>Testing</td>
<td>12,592</td>
<td>9,344</td>
<td>9,225</td>
</tr>
</tbody>
</table>

Table 3.3: Annotated Concept Counts in i2b2 Challenge Corpus

3.3.2 System Description

In this section, we present an overview of the technologies and features implemented in the development of our initial MER system on the Fourth i2b2/VA challenge dataset.

As discussed in Chapter 2, machine learning approaches are often preferred over rule-based methods due to their high performance, adaptability and less of a requirement for manual effort given its associated expense. As such, we will continue this trend with a key focus on feature development. This leads us to our first decision, choosing the machine learning algorithm to form the basis of the system. The recognition of medical entities may be thought of as a sequence labelling problem. The two most popular and high-performing algorithms for sequence labelling are \textit{Conditional Random Fields (CRFs)} and \textit{Structural Support Vector Machines (SSVM)}.

CRFs are a sequence modelling framework first introduced in Lafferty et al. (2001) that addresses the label bias problem as found in other Markov models. Designed as an undirected graphical model, CRFs model the conditional probability of a possible label sequence $Y$ given an observation label sequence $X$, $P(Y|X)$. For example, in the case of named entity recognition $X$ may represent a sequence of tokens where $Y$ represents a sequence of named entity tags e.g. \{Beginning, Intermediate, Outside\}.

The SSVM presents a feasible alternative to the CRF. However, before presenting SSVMs we must first discuss the SVM as a classification tool. The SVM (Cortes and Vapnik, 1995) represents each instance as an N-dimensional vector, where N is the number of features. The key idea of the SVM is to identify a hyperplane that separates vectors in the data set into two classes such that the distance between the nearest point from each class to the hyper-
plane is a maximum. This distance is known as the maximum margin. In the case of an SSVM, the formulation is changed such that the SVM predicts the output label $Y$ given the source $X$.

The literature has shown that while there are suggestions that the SSVM may be more generalizable (Nguyen and Guo, 2007), this is far from conclusive (Keerthi and Sundararaja, 2007). However, more concrete is the fact that while CRF training is intractable, SSVM computation can be computed in linear time through the use of a cutting-plane method (Joachims et al., 2009). It is for this tractable training and high-performance that our MER system will be built upon a Structural SVM framework, rather than commonly used CRFs. The use of Structural SVMs requires optimization on several fronts, namely kernel choice and the cost parameter $C$. In the case of non-linear classification, alternate kernels may be used such as a polynomial or Radial Base Function (RBF) kernel. The $C$ value determines the size of the margin. A small $C$ value searches for the largest possible margin for the hyperplane, increasing the number of misclassified instances. A larger $C$ value indicates to select a smaller margin thus minimizing misclassifications. To perform this optimisation, the system was evaluated through 10-fold cross validation to arrive at the decision to use a linear kernel, with an optimised $C$ parameter of 20,000.

To treat MER as a sequence labelling problem, we must define a set of possible labels, or tags. Traditionally tags are produced in the BIO tagging format, such that B marks the beginning of an entity, I signifies a token following B that is part of the same concept. O is used to mark tokens outside of a medical entity span, as shown in Figure 3.2.

<table>
<thead>
<tr>
<th>Patient presented with heart palpitations.</th>
</tr>
</thead>
<tbody>
<tr>
<td>O O O B I</td>
</tr>
</tbody>
</table>

Figure 3.2: Example of BIO tagging format

More recently the use of a BIESO tagging format has been posited in the clinical domain (Tang et al., 2012). BIESO adds two new tag types, E - which marks the end of an entity and S, which marks an entity composed of a single token. This tagging structure aims to create a finer granularity in deciding the function of a token. For example, to detect tokens that localise a medical condition, e.g. “left lateral eye deviation” rather than simply “eye deviation”. Similarly, the use of a single-token tag allows for instances of concepts that contain a single token. As some concepts may occur more frequently in term dependencies
e.g. cancer is often preceded by a location as in breast cancer or bowel cancer, the BIO format may overgeneralise this to recognise any noun preceding the term cancer as a noun compound. A single-token tag eliminates such occurrences.

Results in the literature have shown superior performance of BIESO tagging over the BIO format in certain instances without discussion of why this is the case. We will first investigate the impact of BIESO tagging on our dataset in Section 3.3.3 to identify its strengths and weaknesses with respect to the more traditional BIO format, before deciding which format is best suited to the task of Medical Entity Recognition.

3.3.2.1 Feature-Sets

In this section, we present a breakdown of the features used in performing Medical Entity Recognition on the i2b2/VA dataset. The features described in this section are based on those discussed in Section 2.4.

- **Lexical Features**: Lexical Features analyse the make-up of of the token being investigated, such as checks for capitalisation. While often used in MER approaches in the literature, there are no strict set of lexical features for entity recognition. A full breakdown of the individual features that make up this feature set are outlined in Table 3.4.

- **Affix Features**: Affix features are created from the prefixes and suffixes of the current token. Affixes of size 2, 3 and 4 are taken to generate features. These features are encoded in binary form, where 1 indicates the occurrence of an affix, with 0 indicating its absence. For example, with a token tracheotomy the features suffix_4_tomy and and prefix_4_trac are both set to 1. The use of these affix features generalises the token, relating it to body parts, as in “trac-” as well as treatments in “-tomy”.

- **Part-Of-Speech Features**: Part-Of-Speech features indicate whether a token is belongs to a certain tag in the Penn Treebank (Marcus et al., 1993). These features are also encoded in binary form with 1 indicating it belongs to a given tag and 0 that is does not. For example, given the token “Hordeolum” which is tagged as NN, the feature NN is set to 1 for the current token. Windowing is used for this feature set, in that for each token we also extract token features within a window $W$ surrounding the current token. For experimentation, we have chosen $W$ equal to 3, 5
<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>isBodyPart</td>
<td>A binary feature to check if a token is body part, e.g. <em>leg</em> localises the condition <em>leg edema</em>, thus would be part of an entity span.</td>
</tr>
<tr>
<td>isDeterminer</td>
<td>A binary feature to check if a token is a determiner, e.g. <em>the</em>, <em>a</em> and other determiners should not be tagged as part of an entity span.</td>
</tr>
<tr>
<td>isDisjunct</td>
<td>A binary feature to check if a token is a disjunctive particle. Disjunctions may separate entity spans.</td>
</tr>
<tr>
<td>isConjunct</td>
<td>A binary feature to check if a token is a conjunction. Conjuncts may separate entity spans, e.g. &quot;<em>vomiting and migraines</em>&quot; should be recognised as two entities.</td>
</tr>
<tr>
<td>isSymbol</td>
<td>Binary check of a token is a symbol e.g. &quot;+&quot;, &quot;-&quot;. This is for shorthand + = positive, - = negative, and are often followed by a test result.</td>
</tr>
<tr>
<td>isAcronym</td>
<td>Binary check if string is formatted as an acronym. Entities that are multi-word or complex, frequently are transcribed as an acronym.</td>
</tr>
<tr>
<td>hasOneCapital</td>
<td>Binary check of token if it contains a capital letter.</td>
</tr>
<tr>
<td>moreThanOneCapital</td>
<td>Binary check of a token containing more than one capital letter. A reinforcing feature for discovering acronyms.</td>
</tr>
<tr>
<td>allcaps</td>
<td>A binary check that token is fully capitalised. Fully capitalised acronyms frequently are concepts, with lowercase acronyms not expressing concepts. For example, <em>prn</em> is not a medical entity but Coronary Artery Disease, a medical problem, is more frequently expressed as <em>CAD</em> than <em>cad</em>.</td>
</tr>
</tbody>
</table>

Table 3.4: Lexical Features used in UCD_MER System
and 10. Each feature indicates whether it occurs to the left, or right to the current token

- **Token Features:** Token features provide an indicator of a token’s normalised form, as generated by cTakes. These are binary features. Every token in the corpus is a feature, with a feature set to 1 if the token is of the word form for that feature. e.g. fractures, fracture both set the fracture feature to 1. Similar to Part-Of-Speech features windowing is also performed. Windows are implemented in the same fashion as the Part-Of-Speech features with $W$ equal to 3, 5 and 10 with an indicator signifying it is to the left or right of the current token.

- **Clustering Features:** Token clustering provides a means for the machine to infer related words, based on their distributional contexts. An example of this may be drug names. For example, these frequently occur in similar contexts such as within lists or in certain patterns e.g. aspirin for headache. To leverage this information, as in other works (Tang et al., 2012), the Brown clustering algorithm (Brown et al., 1992) is used. In using this algorithm, hierarchical clusters are created and the path to each cluster is represented as a bit string. For example, Figure 3.3 shows a toy example of hierarchical clustering. In this example, terms $t_M$ through $t_N$ belong to the cluster with the path of ‘000’ whereas terms $t_O$ through $t_P$ belong to the cluster with the path bit string ‘001’. In our clustering we found that this string of bits amounted to twelve. Thus twelve clustering features were created with each feature taking the value of the respective bit in the string.

- **Semantic Features:** While surface and shallow features do add to the MER task through generalising patterns and sequences found in texts, deeper features such as that of clustering and further semantic information may also aid the task as shown in the literature (de Bruijn et al., 2010). To leverage semantic information, concept unique identifiers (CUIs) from the UMLS Metathesaurus are generated through cTakes to normalise the semantic content of a token. This feature is also binary, with all possible CUIs as features, the presence of a given CUI marks a feature as 1, otherwise 0.

In using all of the feature types described, the number of features for each instance amounts to over half a million.
Overview of System Workflow

Figure 3.4 shows the overall workflow of constituent components in creating the UCD_MER system. Though not required on the i2b2 dataset, a sentence splitting module is included. This module leverages classes and medical text models from the LingPipe toolkit \(^1\) in order to accurately split sentences in clinical reports when required. The second stage of processing performs analysis on the texts with cTakes \(^2\) and the Brown clustering algorithm \(^3\). The third stage uses the document collection, as well as the output of stage two in order to extract the necessary features. Finally, training and classification of tokens is performed using the SSVM implementation svm\(^{hmm}\) \(^4\).

3.3.3 Evaluation Results

In this section, we present the evaluation of our MER system on the subset of the Fourth i2b2/VA challenge dataset.

As previously discussed in Section 3.3.2, although CRFs are commonly applied to NER tasks, its intractable training times has led the research community to seek a more efficient approach without sacrificing its accuracy. The SSVM is

---

\(^1\)www.alias-i.com
\(^2\)http://ctakes.apache.org/
\(^3\)https://github.com/percyliang/brown-cluster
\(^4\)http://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html
one such approach, achieving state-of-the-art performance while performing training in linear time. For our experimentation an SSVM is used. The system uses a linear kernel and through 10-fold cross-validation we have optimized the c-value to be 20,000.

Typically systems apply the BIO tagging format, however, state of the art approaches utilise the BIESO tagging format (Tang et al., 2012). Both of these approaches have been discussed in Section 3.3.2. To identify the tagging format best suited to the problem, identical systems were built, differing only in the tagging format used. The results of these initial tests are displayed in Table 3.5.

<table>
<thead>
<tr>
<th>Tagging</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIESO</td>
<td>0.770</td>
<td>0.741‡</td>
<td>0.756</td>
</tr>
<tr>
<td>BIO</td>
<td>0.776</td>
<td>0.727</td>
<td>0.751</td>
</tr>
</tbody>
</table>

Table 3.5: UCD_MER: BIO Vs. BIESO Tagging on the i2b2 Corpus. ‡ indicates a statistically significant difference in performance than the next ranked system.

The results of both systems are largely similar, however we note statistical significant improvement of recall in the use of BIESO tagging, owing to its focused approach towards single token entities. Recall for these systems is key in an IR setting. While precision is important, were a system in this setting too focused on precision, it may miss pivotal entities in a text thus deeming it incorrectly irrelevant to a given information need. For this reason, we will

---

5Significance is calculated using the Paired Student’s t-test with a p-value = 0.05, for all experiments in this chapter.
apply BIESO tagging throughout the entity detection task. This provides an answer to the research question *What tagging format best suits the recognition of continuous entities*, in that **BIESO tagging facilitates higher recall in the detection of entities**.

The results of the MER system using BIESO tagging and all feature sets with affix size and window size set to 4 and 10, respectively are presented in Table 3.6.

<table>
<thead>
<tr>
<th>Evaluation Type</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact</td>
<td>0.770</td>
<td>0.741</td>
<td>0.756</td>
</tr>
<tr>
<td>Inexact</td>
<td>0.878</td>
<td>0.823</td>
<td>0.850</td>
</tr>
</tbody>
</table>

Table 3.6: Exact and Inexact evaluation of UCD_MER using all features

We see favourable performance of the system, with exact and inexact f-scores of 0.756 and 0.850, respectively. These results indicate that the system is quite effective in determining the occurrence of an entity in the text, however its ability in defining concept spans often differ to the gold standard. A breakdown of performance per entity type is given in Table 3.7. Notably, the highest performance is in recognising tests. The recognition of medical problems is the lowest among the three entity types. A contribution to this is the variation of problem and test occurrences. Instances of class *test* are likely to be found under similar headings such as *HEENT*, *ECHO* or *Hospital Course*. Secondly, they often are preceded by determiners, while followed by a verb, for example “His bloods show...”. This is in contrast to medical problems, that are less generalisable. They may occur under any heading in the text, and while there are generalisable patterns, not all of them have been implemented due to the risk of over-fitting.

<table>
<thead>
<tr>
<th>Entity Type</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.770</td>
<td>0.741</td>
<td>0.756</td>
</tr>
<tr>
<td>Problem</td>
<td>0.747</td>
<td>0.728</td>
<td>0.737</td>
</tr>
<tr>
<td>Treatment</td>
<td>0.782</td>
<td>0.712</td>
<td>0.745</td>
</tr>
<tr>
<td>Test</td>
<td>0.791</td>
<td>0.789</td>
<td>0.790</td>
</tr>
</tbody>
</table>

Table 3.7: Breakdown of UCD_MER performance per entity type

**Creating a comparable system**

The dataset used in these experiments is a subset of that used in the Fourth i2b2/VA challenge for Concept Identification. As the full dataset is not available, this hinders comparative analysis between the results presented in this thesis and in i2b2 literature. To facilitate further comparative research in the area, we have evaluated our approach with respect to two systems reported
Figure 3.5: UCD_MER Versus Stanford NER and ABNER (Exact Evaluation)

by Kang et al. (2010), namely the Stanford Named Entity Recognizer (Finkel et al., 2005) and ABNER (Settles, 2004), with performance noted on the full i2b2 dataset. Both of these systems are free to download. The systems were trained on the i2b2 subset training data, and evaluated on the test data, allowing for comparison between system results at the i2b2 challenge and in this thesis.

Figure 3.5 shows the results of the Stanford NER system and ABNER on the i2b2 subset and the full set as reported in Kang et al. (2010). The results are calculated using the exact matching criteria and are shown in respect to the MER system developed in this thesis, UCD_MER. Performance for both ABNER and Stanford remains largely similar across the full-set and subset. A point of note is the drop in recall for both approaches when applied to the full test set, highlighting their focus on precision. UCD_MER posts the best recall scores, a feature of the IR focused system. However, differences between systems are statistically insignificant, with a difference of 2% separating the f-scores across all three systems.

A deeper error analysis of the systems further showcased the importance of semantic knowledge in medical entity recognition systems. As discussed in the previous chapter, UCD_MER utilises the UMLS Metathesaurus to normalize semantic information from raw text, where the Stanford and ABNER systems do not. This is particularly advantageous when dealing with drug names e.g.
senokot, plavix as well as localizing medical conditions, e.g. ABNER and Stanford NER both incorrectly recognized ulcer rather than the correct span ulcer on his lip. However, this level of information also had its disadvantages. With a focus on semantic information in the UCD_MER system, modifications of concepts were poorly recognized e.g. identifying warfarin rather than the correct span 10mg warfarin. This was particularly found to be the case where concepts are featured in the first or last two tokens of a sentence.

The length of concepts greatly impacted all systems. The mean number of tokens and number of characters in a concept are 1.1 and 10.3, respectively. All systems encountered difficulty in recognising terms typically with lengths less than 4 characters, in particular acronyms such as dvt, jvd and dm-2. At the other end of the scale, systems had difficulty with exceptionally long tokens e.g. 8mm ulcer at junction of distal descending colon and sigmoid colon, however spans of such length were a rarity among the corpus.

UCD_MER detected more false positives than any other system due to its focus on recall. This owes to the use of semantic information that generalises the textual representation of entities. Despite this, the type of false positives produced by all three systems were largely similar. For example, it is often the case that a body part followed by text signifies an attribute in the patient being tested, e.g. ABDOMEN: NONTENDER. However, all systems often, and incorrectly extrapolated that the single sequence ABDOMEN: was in fact the test, rather than the entire sequence. Misclassification across concept types was a relatively rare occurrence. One occurrence incorrectly detected is intracerebral haemorrhage drainage as a treatment rather than a problem. The attachment of the term drainage was the cause of such issues as the system detected its occurrence as the drainage of a wound, which was typically the case in the training data.

The strict nature of the metrics meant that the exclusion of a non-key term in a concept, such as a determiner at the head of a concept, registered a false positive/negative for the system. To discount such harsh criteria, a relaxed metric is also used where the spans may only be overlapping but still require the correct concept type. The results of this evaluation are displayed in Table 3.8.

Through relaxing the matching criteria, we see a significant improvement in scores across all systems. Intuitively, the differences between systems are non-significant, highlighted by the minute difference in f-score. While the systems provide competitive results, there is still room for improvement. To further improve the performance of UCD_MER, we must investigate the contribution...
Table 3.8: Inexact evaluation of UCD_MER, ABNER and Stanford NER on i2b2 Challenge Corpus

<table>
<thead>
<tr>
<th>System</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stanford NER</td>
<td>0.915</td>
<td>0.823</td>
<td>0.855</td>
</tr>
<tr>
<td>ABNER</td>
<td>0.901</td>
<td>0.814</td>
<td>0.853</td>
</tr>
<tr>
<td>UCD_MER</td>
<td>0.878</td>
<td>0.823</td>
<td>0.850</td>
</tr>
</tbody>
</table>

Figure 3.6: UCD_MER Feature Performance on i2b2 Challenge Subset

What are the contributions of each of the feature sets?

In Figure 3.6 we see the precision, recall and f-score for individual feature-sets in the i2b2 medical entity task using the exact matching criteria. Lexical features perform poorly on their own, with all scores less than 0.1. This is not to say they are entirely redundant, as they may capture more specific instances. However, they generalise poorly due to the limited set of rules. For example, features such as capitalisation may provide an indicator to acronyms such as MRSA, AIDS, however given that this is only one of many different types of concepts found in the corpus, lexical features do not suffice on their own.

Similarly, clustering and semantic features performed poorly on their own. The performance of semantic features may be due to the over-generalising of concept identifiers. cTakes was used to extract CUIs for each token. As multiple CUIs are generated without restriction this could harm the effectiveness of the system. A more precise method of identifying CUIs, such as restrict-
ing to particular semantic types may aid performance. While clustering has been shown in the literature to aid the MER task in a variety of domains in combination with other feature types, they are far from effective on their own. The Brown clustering algorithm clusters tokens based on bigram statistics, as a result missing the wider context of tokens. The restrictive nature of the clustering analysis is the root of the marked improvement in precision.

Part-of-speech information showed improvements in precision over other feature types. However, f-scores are still reduced due to low recall. However, as the window size increases we see an increase in precision. As the mean sentence length is 19.3 words, the window size is capped at 10 to ensure maximum feature impact.

Token features contribute the most to the task, achieving an f-score of 0.73. While this may indicate that the word form of concepts are frequent, it also highlights the impact of word order prior to, and following the string. For example, medical conditions are frequently preceded by strings such as Patient admitted with, Patient suffering from. Further, instances of class test are frequently described at the head of a sentence fully capitalised followed by a single token, or a value. The choice of window size also highlights the importance of capturing these patterns. The increase in window size correlates with a growth in precision. However, increasing the window size has also seen a drop in recall. The window must be large enough to capture meaningful patterns in the text, while still being restrictive enough so as to not to overgeneralise.

Finally, we investigate the effect of varying affix length. Affixes of size 2 had poor performance, with gains in performance where the affix size is increased. The growth of precision and recall is much more evenly distributed than that of POS and cluster features. Much like the token features, affix features capture patterns among tokens. As we increase the size of affixes, we are capturing much of the information available to token features, thus reinforcing the impact of token features. With the restriction to smaller forms, this allows for more generalisable features, e.g. the suffix -graphy typically applies to tests.

3.3.4 Discussion

In this section, we presented and evaluated a machine-learning approach to Medical Entity Recognition. Our system adapted the Structural SVM rather than the commonly used CRF approach due to the SSVM’s linear complexity training method, this is especially useful as our feature space contains over
half a million features. These features were split into six groupings: lexical, affix, POS, token, cluster and semantic. In our analysis of feature contributions we saw that using only the tokens themselves as features within a certain window created a strong baseline, achieving an f-score of 0.73 on the exact metric. When feature-sets were combined, performance of the UCD_MER system was comparable to the Stanford NER and and ABNER systems.

3.4 CLEF/ShARe Shared Task

In this section, we present a description and evaluation of an MER system developed as part of the CLEF/ShARe shared task. The system follows on from the UCD_MER system described in Section 3.3. However, the system must be adapted due to differences in the task definition. Firstly, while the i2b2/VA challenge featured three possible concept classes, the CLEF/ShARe requires the identification of one, namely disorder. A disorder is more or less semantically equivalent to a problem defined in the i2b2/VA challenge. However, they differ in their structure. Entities in the i2b2/VA challenge must be a continuous sequence of tokens, whereas disorders may be disjoint in the CLEF/ShARe task. That is to say, a disorder span may be broken into segments by tokens that are in fact not part of the concept span. Thirdly, the CLEF/ShARe task introduces concept normalisation, whereby marked spans are to be mapped to concept identifiers in the UMLS Metathesaurus. In Section 3.4.2, we provide an overview of the MER and Concept Normalisation system designed to the task. In Section 3.4.3, we provide a breakdown and evaluation of MER and Concept Normalisation performance.

3.4.1 Dataset Overview

The corpus for CLEF/ShARe eHealth 2013 was made up of clinical narrative reports including discharge summaries, ECG reports, echo reports and radiology reports. These reports were generated in an ICU setting and were obtained from the MIMIC-II database subject to ethical approval. The training and test corpus contained 200 and 100 reports, respectively.

The gold standard annotation of the dataset was generated through the ongoing ShARe (Shared Annotated Resources) project. Unlike the i2b2 dataset, only disorder mentions i.e. problem concept in the i2b2/VA corpus, are annotated.
The annotation of the dataset consists of two stages: (1) The annotation of disorder offsets. (2) Mapping a span to a UMLS CUI. Two professional coders annotated the gold standard dataset followed by an adjudication step.

A disorder is defined as any span of text that can be mapped to a concept that belongs to the Disorder semantic group in the UMLS. The semantic group consists of the following UMLS semantic types:

- Congenital Abnormality
- Acquired Abnormality
- Injury or Poisoning
- Pathologic Function
- Disease or Syndrome
- Mental or Behavioural Dysfunction
- Cell or Molecular Dysfunction
- Experimental Model of Disease
- Anatomical Abnormality
- Neoplastic Process
- Signs and Symptoms

In breaking down the dataset between disjoint and joint disorders in Table 3.9, we see a clear disparity between types of disorder where joint disorders outnumber disjoint disorders by a factor of 10:1. Nonetheless, measures must be taken in order to account for disjoint entities. Therefore, new feature sets will be created to account for disjoint disorders.

<table>
<thead>
<tr>
<th></th>
<th>Disjoint Entities</th>
<th>Continuous Entities</th>
<th>Total Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>651</td>
<td>5165</td>
<td>5816</td>
</tr>
<tr>
<td>Test</td>
<td>439</td>
<td>4912</td>
<td>5351</td>
</tr>
</tbody>
</table>

Table 3.9: Breakdown of CLEF/ShARe Concept Annotations
3.4.2 Systems for MER and Concept Normalization

As stated, the CLEF/ShARe task differs to that of the i2b2/VA challenge in several respects. However, at its core it is a Medical Entity Recognition task. As such, we propose a system that leverages the same core technologies as the UCD_MER system, as shown in Figure 3.7. The system uses the SSVM to perform sequence prediction. 10 fold Cross-Validation experiments led us to choose 20,000 as an optimized C-value.

We maintain that the CLEF/ShARe task is a sequence labelling task. However, as disorders may be disjoint, the previously investigated tagging systems such as BIO and BIESO as we must give special consideration to tokens that occur within a concept span, but are not members of a given concept. Figure 3.8 shows an example of a disjoint disorder in the red text, namely “eczema of the extremities”. However, this disorder is separated by the tokens in blue. The span here could be generalized as “LOC : tok1 tok2 DISORDER” where LOC is the location of the disorder. Taking this pattern into account, we propose a new tagging format, BIESTO (T = beTween), whereby tokens occurring between two linked disjoint entities are tagged T as illustrated by Figure 3.9.

The left atrium is moderately dilated.

Figure 3.9: Example of BIESTO tagging
Furthermore, we adapt the feature-set described in Section 3.3.2.1 to improve performance and create custom features for changes in the CLEF/ShARe task. The first aspect of the feature set change is the inclusion of the affix and lexical features within the token feature set. As a result, instead of inspecting the current token, we build a set of contextual information about tokens within a window of the current token. Secondly, we implement two new lexical features:

- **containsVowels**: This binary feature determines if a token contains vowels, as the UCD_MER system encountered errors in acronyms containing no vowels e.g PTSD.

- **isSlash**: A binary feature identifying if a token is a “/” or “\” symbol. This is again born from i2b2 errors where concepts containing a slash character were incorrectly split in two.

In an attempt to link disjoint entities, we make use of cTakes output in an attempt to link semantic information using the features below:

- **isDisorderCUI**: A binary feature indicating if the CUI for the current token is a member of the disorder semantic group.

- **isBetweenDisorderCUIS**: A binary feature to determine if a token occurs between tokens that have disorder CUIs.

- **matchesAnotherCUI**: A binary feature that checks if the current token’s CUI matches another Disorder CUI in the current sentence.

These features combine with the Structural Support Vector Machine algorithm to create **UCDCSI.1**. **UCDCSI.1** leverages cTakes to form features based on CUIs in the UMLS Metathesaurus.

However, cTakes possesses the ability to recognize medical entities although not entities that are disjoint. To investigate if cTakes is a feasible solution in recognizing disjoint medical entities, we create **UCDCSI.2**, a baseline system to provide a reference point in the need of bespoke solutions. This baseline system uses cTakes and a set of post-processing rules with the aim of recognizing disjoint entities. These simple rules consisted of two checks. If cTakes found a concept span whose CUI fits the disorder semantic type, a check is performed within ten tokens of it to find a matching CUI. If found, the two spans are then linked into a disjoint entity.
Concept Normalization

For concept normalisation, both systems utilised cTakes output to choose a CUI for the annotated span. Two methods were used: A frequency based method, whereby the most common concept among tokens is chosen as the normalised CUI; and an edit-distance based metric where the edit distance between the annotated span and the text of CUIs outputted by cTakes is calculated, with the CUI with the lowest edit distance chosen. The algorithms for both rules are shown in Algorithm 1 and Algorithm 2 below.

Algorithm 1 Choosing a CUI based on its frequency

```
TOKENS ← INPUT
COUNT ← {} 
for all T in TOKENS do 
    possCUIS ← (T→GetCUIS())
    for all cui in possCUIS do
        if (isDisorder(cui) == true) then
            COUNT[cui] += 1
        end if
    end for 
end for
bestCUI ← (COUNTS→GetBestCUI)
return bestCUI
```

Algorithm 2 Choosing a CUI based on edit-distance

```
TOKENS ← INPUT
ALL_CUIS =[]
for all T in TOKENS do
    possCUIS ← (T→GetCUIS())
    for all cui in possCUIS do
        if (isDisorder(cui) == true) then
            ALL_CUIS.add(cui)
        end if
    end for
end for
bestMatch ← 0.0
bestCUI ← []
for all cui in ALL_CUIS do
    currScore ← editDist(TOKENS, cui.getString())
    if (currScore ≥ bestMatch) then
        bestMatch ← currScore
        bestCUI ← cui
    end if
end for
return bestCUI
```
3.4.3 Evaluation Results

In this section, we present the results of a second Medical Entity Recognition system to address the tasks defined as part of the CLEF/ShARe 2013 Challenge. As discussed in Section 3.4.2, the CLEF task differs to the i2b2/VA task in three ways. Firstly, we are only concerned with the recognition of medical problems/disorders. Second, entities may be disjoint i.e. an entity may be a sequence of non-continuous tokens. The final difference is the inclusion of a concept normalisation task. Through concept normalisation we standardise the vocabulary of clinical documents such that level of variation in expressing concepts is reduced. This is a frequent discussion point in the Clinical IR domain, and is therefore an application area warranting investigation.

In Section 3.4.3.1 we present performance in the Medical Entity Recognition task, with Concept Normalisation performance described in Section 3.4.3.2.

3.4.3.1 Disorder Recognition

In MER shared tasks, systems are typically custom-built, domain-specific machine-learning approaches. However, with the availability of off-the-shelf programs and toolkits such as cTakes and MetaMap, we wish to ascertain if it is necessary to create custom-built systems. UCDCSI.2 manipulates cTakes output as described in Section 3.4.2, in order to achieve the required output. This system forms our baseline in evaluating our own machine-learning based system, UCDCSI.1. UCDCSI.1 leverages many of the features designed in the UCD_MER system, with semantic features being more finely honed as well as the combination of lexical features and token features as described in Section 3.4.2. The official evaluations of both systems in the concept identification task for exact and inexact evaluations are presented in Table 3.10 and Table 3.11, respectively.

<table>
<thead>
<tr>
<th>System Name</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCDCSI.1</td>
<td>0.745</td>
<td>0.587</td>
<td>0.656</td>
</tr>
<tr>
<td>UCDCSI.2 (baseline)</td>
<td>0.268</td>
<td>0.175</td>
<td>0.212</td>
</tr>
</tbody>
</table>

Table 3.10: Official Submission Results for CLEF/ShARe Task 1a, Exact Evaluation

UCDCSI.1 significantly outperformed the baseline UCDCSI.2 on both exact and inexact evaluations, reinforcing the opinion that custom built systems are required but also indicating shortcomings in post-processing rules applied to
<table>
<thead>
<tr>
<th>System Name</th>
<th>System Name</th>
<th>0.922</th>
<th>0.758</th>
<th>0.832</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCDCSI.1</td>
<td>UCDCSI.2 (baseline)</td>
<td>0.512</td>
<td>0.339</td>
<td>0.408</td>
</tr>
</tbody>
</table>

Table 3.11: Official Submission Results for CLEF/ShARe Task 1a, Inexact Evaluation

Table 3.12: Disjoint and Joint F-scores for author’s submissions CLEF/ShARe Task 1a, Exact Evaluation

The baseline approach recognised none of the 438 disjoint disorders in the test set. However as shown by its performance on joint entities, post-processing rules were not necessarily the root of the baselines issues. Although the systems are evaluated on the exact criteria, we see that when applying inexact criteria despite almost a doubling in f-score, UCDCSI.2 was no match for UCDCSI.1. This provides a strong indicator that a machine-learning approach which leverages deep semantic information as well surface based features are best suited to the task.

UCDCSI.1 posts competitive results, ranked within the top ten performing systems at the challenge for exact and inexact matching as shown in Tables 3.13 and 3.14, respectively. At the time of writing, system descriptions are unavailable for participants at the challenge, as well as no breakdowns for joint and disjoint spans. In spite of this, we may proceed with an analysis of the UCDCSI.1 system.

**Analysing the errors produced by the UCDCSI.1 run.**

Much like our analysis in Section 3.3.3, exact matching criteria often proves too strict for the evaluation of MER systems. This is shown in the leap of performance scores between Table 3.13 and Table 3.14 once the criteria is relaxed. This indicates that systems typically correctly identify spans within the document, but there are often inconsistencies in the specificity of gold standard annotations that tend to harm performance across systems.

In report 17644-017974-DISCHARGE_SUMMARY the precision of the system in identifying the most descriptive span was disruptive to its overall score. For example, the system identifies the spans right colon adenocarcinoma, recurrent cancer rather than the more general, and correct, colon adenocarcinoma, can-
<table>
<thead>
<tr>
<th>System Name</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTHHealth_CCB.2</td>
<td>0.800</td>
<td>0.706</td>
<td>0.750</td>
</tr>
<tr>
<td>UTHHealth_CCB.1</td>
<td>0.831</td>
<td>0.663</td>
<td>0.737</td>
</tr>
<tr>
<td>TeamNCBI.1</td>
<td>0.768</td>
<td>0.654</td>
<td>0.707</td>
</tr>
<tr>
<td>TeamNCBI.2</td>
<td>0.757</td>
<td>0.658</td>
<td>0.704</td>
</tr>
<tr>
<td>TeamCLEAR.2</td>
<td>0.764</td>
<td>0.624</td>
<td>0.687</td>
</tr>
<tr>
<td>TeamMayo.1</td>
<td>0.800</td>
<td>0.573</td>
<td>0.668</td>
</tr>
<tr>
<td>TeamCLEAR.1</td>
<td>0.755</td>
<td>0.573</td>
<td>0.656</td>
</tr>
<tr>
<td>TeamMayo.2</td>
<td>0.697</td>
<td>0.574</td>
<td>0.629</td>
</tr>
<tr>
<td>TeamCORAL.2</td>
<td>0.796</td>
<td>0.487</td>
<td>0.604</td>
</tr>
</tbody>
</table>

Table 3.13: Top 10 Performing Systems at CLEF/ShARe Task 1a, Exact Evaluation

<table>
<thead>
<tr>
<th>System Name</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTHHealth_CCB.2</td>
<td>0.925</td>
<td>0.827</td>
<td>0.873</td>
</tr>
<tr>
<td>UTHHealth_CCB.1</td>
<td>0.954</td>
<td>0.774</td>
<td>0.854</td>
</tr>
<tr>
<td>TeamNCBI.2</td>
<td>0.904</td>
<td>0.805</td>
<td>0.852</td>
</tr>
<tr>
<td>TeamNCBI.1</td>
<td>0.910</td>
<td>0.796</td>
<td>0.849</td>
</tr>
<tr>
<td>TeamMayo.2</td>
<td>0.939</td>
<td>0.766</td>
<td>0.844</td>
</tr>
<tr>
<td>TeamCLEAR.2</td>
<td>0.929</td>
<td>0.759</td>
<td>0.836</td>
</tr>
<tr>
<td>TeamAEHRC.2</td>
<td>0.886</td>
<td>0.785</td>
<td>0.833</td>
</tr>
<tr>
<td>UCDCSI.1</td>
<td>0.922</td>
<td>0.758</td>
<td>0.832</td>
</tr>
<tr>
<td>TeamCLEAR.1</td>
<td>0.937</td>
<td>0.705</td>
<td>0.804</td>
</tr>
<tr>
<td>TeamArvindWVU.1</td>
<td>0.788</td>
<td>0.814</td>
<td>0.801</td>
</tr>
</tbody>
</table>

Table 3.14: Top 10 Performing Systems at CLEF/ShARe Task 1a, Inexact Evaluation
cer. As this is a common trait among errors located using the exact matching criteria, we will focus on relaxed criteria errors.

Negation is a key factor in the system generating false negatives. For example, the system fails to recognise No MVP, No MS. Extending this finding, the system also tends to miss acronyms with two or less characters. In particular, these errors occur in sentences where abbreviations are frequent and medical shorthand is common, e.g. V- Irregular, +s1/s2, no s3/s4. II/VII late diastolic murmur, where s3 is the entity to be identified.

A third cause of false negatives is the use of comma delimited lists. For example, in the following list taken from a report, history of headaches, hypertension, diabetes, hepatitis C, depression, it is common for the system to miss any of hypertension, diabetes and hepatitis C. There are two factors at play which cause the misclassification of the mid-elements of the list. Firstly, history of CONDITION is a common pattern used to identify conditions in the text. Comma delimited lists are frequently used to list symptoms, or the the results of an inspection/exam, terminating in the author positing a possible diagnosis.

As shown by the high precision of the system, false positives were a rare occurrence in comparison to false negatives, with 0.2 false positives for every false negative. In spite of this, there are some identifiable trends among the errors. As CUIs are attached to individual tokens, rather than larger phrases it results in the misclassification of tokens. For example, as cTakes extracts multiple CUIs per token where available, the token pressure may be mapped to the concepts pressure ulcer or intracranial pressure. Both of these concepts, are within the disorder semantic types and so switches value of the feature that signifies a token is a disorder, from a 0 to 1. However, if we take it that it is a member of the bigram blood pressure, this is clearly incorrect.

The generalisation power of affixes also has a role to play in the generation of false positives. For example, given the phrase hip arthroplasty, using a prefix of size three we may arrive at the following hip, art. This may then incorrectly generalise to the concept hip art(hritis). Token features are also susceptible to over generalising. For example, the token painful in the context of an action such as walking is often indicative of a medical disorder. However, this generalisation has proven to cause several false positives where responds to painful stimuli is a test rather than a medical condition.

What are the feature contributions to the task?

Through our error analysis we may identify instances where certain features
may succeed or fail. To determine the exact contributions of these features we may inspect their performance individually as shown in Figure 3.10 and Figure 3.11. The feature selection experiments here differ to those on the i2b2 dataset in that, following our analysis of lexical and affix types, it was found that they could prove useful in identifying generalised patterns in the text. As a result both of these feature types are now combined in the token feature type as explained in Section 3.4.2.

As a result of combining lexical, affix and token features under the token umbrella, we see this feature set outperform all others across all evaluation metrics as in the i2b2 experiments. A key difference between the experimental results however, is the importance of semantic features. This is due to several factors. Firstly, the aim of the task has been slightly altered to require and produce more semantic
The identification of disorder spans aims to facilitate their normalisation, and as such the semantic content plays a role in these spans. What’s more, is the new allowance of disjoint entities requires a means to produce a link between tokens in the text. While syntactic rules may also be used to link disjoint entities, the use of semantic information facilitates the underlying task of normalisation. Thirdly, and perhaps most importantly, the range of semantic types allowed is far more restrictive thus providing a far more effective means of conveying semantic information. For the i2b2 experiments, all types were used across three different entity types which often led to confusion in the system. Through restricting the allowable semantic types, the possibility of over-generalisation is reduced.

Once again we identify poor performance of clustering features and part-of-speech features with a small window. Both features have an emphasis on precision with poor recall indicating little generalisation ability. Despite their poor performance on small windows, f-score growth correlates with an increase of window size for the part-of-speech feature-set. It must be stated that while clustering features perform poorly on their own (Tkachenko and Simanovsky, 2012), they have been shown to tip competitive systems into the state-of-the-art.

Our findings through our feature analysis lead us to explore a more optimal feature set. By optimal, we mean a feature set that has removed redundant feature sets that either negatively impact the system or have no impact at all. By doing so, we aim to reduce computation time while seeing improvements in performance.

Table 3.15: Feature Selection for MER for CLEF/ShARe task, F-score Results.

<table>
<thead>
<tr>
<th>System Name</th>
<th>Exact</th>
<th>Inexact</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>tok_5_sem_clus</td>
<td>0.744</td>
<td>0.6</td>
</tr>
<tr>
<td>tok_5_sem_pos_10‡</td>
<td>0.741</td>
<td>0.599</td>
</tr>
<tr>
<td>tok_5_sem‡</td>
<td>0.734</td>
<td>0.597</td>
</tr>
<tr>
<td>UCDCSI.1‡</td>
<td>0.745</td>
<td>0.587</td>
</tr>
<tr>
<td>tok_10_sem</td>
<td>0.743</td>
<td>0.569</td>
</tr>
</tbody>
</table>

† indicates the best system by a statistically significant margin. ‡ indicates a statistically significant difference in performance than the next ranked system.

Table 3.15 shows a select number of runs generated through feature selection. The additional runs are broken down as follows:

- tok_5_sem: Combines top performing token feature set, with the re-

---

6Feature selection is performed using greedy backward subset search.
moval of affix features, with a window size set to five, along with the second best semantic feature.

- **tok_10_sem**: Same as **tok_5_sem**, with window increased to ten.
- **tok_5_sem_pos_10**: Same as **tok_5_sem**, add pos features with a window of ten.
- **tok_5_sem_clus**: Same as **tok_5_sem**, combines cluster features to investigate their impact.

Three of the four additional runs outperform our initial system UCDCSI.1. The only run it outperformed, **tok_10_sem** indicates that while increasing the window size improves precision, with it achieving the highest inexact precision, it significantly hampers recall. As a result of poor recall, the overall system performance drops. *The top ranking runs all feature a window of size 5, the removal of affix features and the inclusion of semantic information.* While precision of the system drops through the reduction of the feature-sets, the more generalisable feature-sets see a significant improvement in recall and thus f-score. **tok_5_sem_clus** is the top performing run, showing that *while clustering information is not effective solely on its own it does contribute to the task in gaining top performance when combined with other features.*

### 3.4.3.2 Concept Normalisation

In this section, we present and discuss the results of evaluating the Concept Normalisation approaches described in Section 3.4.2. For this task, the system must correctly identify a disorder span and then map it to a CUI in the UMLS Metathesaurus. Two systems are presented, a machine learning approach and a system comprising of post-processing rules for cTakes output. Both systems rely on cTakes to map text to CUIs. However, as cTakes presents numerous possible CUIs, it is the task for the systems to ascertain the most suitable CUI.

The official evaluation results are presented in Table 3.16. For the exact evaluation a true positive is marked as having the correct span as well as the correct CUI. For the inexact evaluation, the span matching criteria is relaxed to count overlapping spans as true positives. Both systems use the frequency based method of selecting a CUI as described in Section 3.4.2.

---

7Significance is calculated using the Paired Student’s t-test with a p-value = 0.05, for all experiments in this chapter
UCDCSI.1 posted moderate results, while UCDCSI.2 performance was disappointing. While the systems use the same set of rules for generating output, there are two key deficiencies of UCDCSI.2. Firstly, the poor span identification of UCDCSI.2 as described in Section 3.4.3.1 has a knock-on effect here, as only correctly identified spans are eligible for concept mapping. This poor span identification also works to its disadvantage as more tokens are introduced to a span, so too does the number of noisy concepts generated.

The marked improvement of UCDCSI.1 over UCDCSI.2 can be attributed to its superior performance in identifying concept spans. However, in spite of this raise in performance it fails to achieve a score within the top 10, as listed in Table 3.17 and Table 3.18.

<table>
<thead>
<tr>
<th>System Name</th>
<th>Accuracy (Exact)</th>
<th>Accuracy (Inexact)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCDCSI.1</td>
<td>0.299</td>
<td>0.509</td>
</tr>
<tr>
<td>UCDCSI.2 (baseline)</td>
<td>0.006</td>
<td>0.035</td>
</tr>
</tbody>
</table>

Table 3.16: Official Author Submission Results for CLEF/ShARe Task 1b: Concept Normalisation

<table>
<thead>
<tr>
<th>System Name</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TeamNCBI.2</td>
<td>0.589</td>
</tr>
<tr>
<td>TeamNCBI.1</td>
<td>0.587</td>
</tr>
<tr>
<td>TeamMayo.2</td>
<td>0.546</td>
</tr>
<tr>
<td>UTHHealth_CCB.1</td>
<td>0.514</td>
</tr>
<tr>
<td>UTHHealth_CCB.2</td>
<td>0.506</td>
</tr>
<tr>
<td>TeamMayo.1</td>
<td>0.502</td>
</tr>
<tr>
<td>KPSCMI.1</td>
<td>0.443</td>
</tr>
<tr>
<td>TeamCLEAR.2</td>
<td>0.440</td>
</tr>
<tr>
<td>TeamCLEAR.1</td>
<td>0.409</td>
</tr>
</tbody>
</table>

Table 3.17: Top 10 Performing Systems for CLEF/ShARe Task 1b, Exact Evaluation

Given the improvement of UCDCSI.1 over UCDCSI.2 due to more accurate span detection, we may investigate the effects of applying our top performing feature-sets as discovered in the previous section. The results of these experiments are set out below in Table 3.19.

The rankings of the concept normalisation mirror that of the disorder recognition task, indicating that as the performance of the span matching goes up, so too will the normalisation performance. Despite improvements, there is still a marked difference in the performance of our top performing normalisation system, and the state-of-the-art, while not much separates them in terms of disorder re-
Table 3.18: Top 10 Performing Systems for CLEF/ShARe Task 1b, Inexact Evaluation

<table>
<thead>
<tr>
<th>System Name</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TeamAEHRC.1</td>
<td>0.939</td>
</tr>
<tr>
<td>TeamNCBI.1</td>
<td>0.897</td>
</tr>
<tr>
<td>TeamNCBI.2</td>
<td>0.895</td>
</tr>
<tr>
<td>TeamMayo.1</td>
<td>0.870</td>
</tr>
<tr>
<td>KPSCMI.1</td>
<td>0.865</td>
</tr>
<tr>
<td>TeamMayo.2</td>
<td>0.860</td>
</tr>
<tr>
<td>UTHHealth_CCB.1</td>
<td>0.728</td>
</tr>
<tr>
<td>UTHHealth_CCB.2</td>
<td>0.717</td>
</tr>
<tr>
<td>TeamCLEAR.1</td>
<td>0.713</td>
</tr>
<tr>
<td>TeamCLEAR.2</td>
<td>0.704</td>
</tr>
</tbody>
</table>

Table 3.19: Effects of Feature Selection on Concept Normalisation

cognition performance. Therefore, improvements must be made in choosing a CUI for a given concept.

In systems to this point, the CUI has been selected based on its frequency among tokens in the concept, with the CUIs extracted via cTakes. This is quite a rudimentary statistic to use in choosing the correct CUI. To expand on this, we use an edit-distance metric as described in Section 3.4.2 to choose the correct CUI. The results of the experimentation are shown in Table 3.20. The results indicate that the choice of metric in selecting the CUI has very little impact across the systems.

Table 3.20: Effects of Edit-Distance CUI Selection on Concept Normalisation

<table>
<thead>
<tr>
<th>System Name</th>
<th>Accuracy (Exact)</th>
<th>Accuracy (Inexact)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tok_5_sem_clus</td>
<td>0.307</td>
<td>0.512</td>
</tr>
<tr>
<td>tok_5_sem_pos_10</td>
<td>0.306</td>
<td>0.510</td>
</tr>
<tr>
<td>tok_5_sem</td>
<td>0.306</td>
<td>0.512</td>
</tr>
<tr>
<td>UCDCSI.1</td>
<td>0.299</td>
<td>0.509</td>
</tr>
<tr>
<td>tok_10_sem</td>
<td>0.299</td>
<td>0.526</td>
</tr>
</tbody>
</table>

Given the little impact of the algorithms used in selecting the CUI from the list of cTakes candidates. We must investigate the initial list outputted by cTakes. On inspection of cTakes output, the system typically outputs several CUIs, of which only one CUI belongs to the required disorder semantic group. As a
result, the choice of algorithm to choose a CUI from the list does not have an
effect on performance. A reason for this is the reliance of methods on cTakes
span detection, which has been determined to be poor. In relying on cTakes
spans, it may incorrectly extract CUIs. This highlights the requirement of cus-
tom built tools for a domain application, such as concept normalisation. Similar to
disorder recognition, while cTakes provides a helpful indicator in guiding a
system, its performance is not to be relied upon for application areas such as
Medical Entity Recognition and Concept Normalisation.

3.4.4 Discussion

In this section, we presented an experimental evaluation of the Disorder Re-
cognition and Concept Normalisation systems on the CLEF/ShARe dataset.
For the Disorder Recognition task we introduced a new tagging format, BI-
ESTO, which aims to tag tokens that occur between disjoint entities. We further
introduce new semantic features developed to link disjoint entities. Feature
analysis has shown token features to be the most effective, with the introduc-
tion of affix and lexical information in the token window function allowing the
generalisations of patterns. Given the requirement for more semantic inform-
ation in the task, we see top performing approaches using both semantic and
clustering feature sets.

3.5 Summary

This chapter presented experimental results of Medical Entity Recognition and
Concept Normalisation systems on the i2b2/VA and CLEF/ShARe corpora.
In relation to our research questions described at the outset of this chapter, we
address the following:

What features contribute the most to Medical Entity Recognition perform-
ance?

Token features have shown the greatest to system performance, indicating that
the recognition of concepts may be found in generalisable patterns in the text.
The effectiveness of using tokens as features rely on the choice of window size.
Our experimentation shows that using a larger window improves precision,
though at the cost of recall. As affix and lexical features contributed little to
the task solely on their own, they are included as part of a token based win-
dow in order to create more generalisable patterns. POS information is also shown to be an effective feature, with a focus on precision. Its usage experiences the same trade-off as the token feature set with a larger window proving more restrictive. The inclusion of semantic and cluster information provided an improvement in system performance, corroborating findings in the literature. However, semantic information is shown to be most effective when semantic types are restricted, as in the CLEF/ShARe experiments. Further, we may see further improvement gains if all the features are generated for all tokens within a window, creating more contextual and pattern information for each instance.

**What tagging format best suits the Medical Entity Recognition task?**

Traditionally, sequence labelling tasks use the standard BIO tagging format in which a given token may be classed as the beginning of an entity, in an entity or outside of an entity. In the case of Medical Entity Recognition, the BIESO format has been put forward in the literature to alleviate issues between concepts that contain many tokens, or just one. In our investigation of both formats, we have discovered that gains are seemingly minimal; however BIESO does see a significant increase in recall over BIO tagging and as such was chosen for our system where recall is key in an IR setting.

**How may we extend Medical Entity Recognition to disjoint entities?**

The recognition of disjoint entities is a new area of research, thus requiring new methods and approaches. In this thesis we investigated extending the BIESO tagging format to also tag text between disjoint entities as well as new features that aimed to connect disjoint entities based on normalised semantic information. Though the approach is promising, there is room for improvement. The inclusion of syntactic dependency information may further improve the recognition of disjoint entities.

**What are the key issues in Concept Normalisation?**

In this chapter, methods leveraging the existing cTakes system are used to normalise identified concepts. Though the system attains an inexact accuracy measure of just over 0.5, our error analysis shows that this is due to reliance on the cTakes system, in particular on its span identification. Due to its relatively poor span identification, cTakes cannot reliably be used to produce accurate CUI candidates according to the CLEF/ShARe evaluations. However, the integration of this system to an IR approach will serve an effective baseline in addressing our thesis question of what is the required performance of NLP
approaches to aid Information Retrieval.

In summary, this chapter has presented experimental findings in the task of Medical Entity Recognition in the context of two shared tasks. We have identified and evaluated machine learning approaches, feature-sets and tagging structures for the Medical Entity Recognition task which will be applied to the Information Retrieval task. Further, we have extended the MER system to perform Concept Normalisation, evaluated on the CLEF/ShARe task dataset. This will serve as an indicator in the Information Retrieval task, as what is the required performance of NLP approaches to aid Information Retrieval.
4.1 Introduction

In this chapter we address the task of Assertion Classification of medical conditions discovered in clinical texts. The assertion status of a medical condition in its most basic form denotes whether the medical condition is currently afflicting the patient described in an electronic medical record. As interest is growing in this research task, its definition is expanding and requiring more exact classifications, such as deciding who experiences a condition and its temporality. In this chapter, we address the following research questions:

- What feature sets contribute most to the task of Assertion Classification?
- How do rule-based systems compare to machine learning based approaches?
- With a multitude of machine-learning algorithms available for classification tasks, which algorithm performs best?
- How can binary classification be extended to a multi-class Assertion Classification task?

In Section 4.3, we present and evaluate machine-learning approaches to resolving two assertion status. Firstly Condition Attribution, where the system decides if a medical condition affects the patient in the report, or some other individual such as a family member. The second Assertion Classification task is named Temporal Grounding. This is again a binary decision task, where classifications decide whether a condition is currently affecting the patient, or was
diagnosed/resolved during a previous admission. The dataset used for this set of evaluations is a subset of the BLULab NLP that was annotated as part of this thesis. In Section 4.4 we expand our approaches for Temporal Grounding and Condition Attribution to six assertion types, changing the task from binary classification to a multi-class classification problem. The system is evaluated on data from the Fourth i2b2/VA Challenge on Assertion Classification, allowing a comparison with other approaches in the research literature. The design and focus of the experiments in this chapter are summarised in Figure 4.1. In the overall Clinical Information Retrieval workflow, Assertion Classification will build on the MER systems and Concept Normalisation systems in Chapter 3, through providing assertion status information about concepts identified and normalised in documents and queries.

Figure 4.1: Diagram of Chapter 4 Systems and Datasets

### 4.2 Evaluation Metrics

To evaluate the performance of Assertion Classification techniques discussed in this Chapter, we use the metrics of precision, recall and f-score as defined in Section 3.2. However, unlike the task of Medical Entity Recognition only exact measures are used as a classification may only be a correct or incorrect classification. Given that Assertion Classification defined in this thesis may be one of six statuses, macro and micro measures will allow us to identify the key gains of the documented approaches.
4.3 UCD Condition Attribution and Temporal Grounding Corpus

In the early stages of this work, the dataset for the Fourth i2b2/VA challenge had not yet been released. For this reason, other means of researching the problem of assertion status were explored. The dataset used in evaluating this pilot system is a subset of BLULab’s NLP repository further described in Section 4.3.1. The system investigates machine-learning approaches to two assertion statuses using a limited set of features. First, where a medical condition mentioned in the clinical report affects another individual, for example a family member. We define this task as Condition Attribution. The second assertion status investigated resolves if a medical condition occurs in the past thus no longer affecting the patient, known as Temporal Grounding. We will outline the experimental setting in Section 4.3.1 and Section 4.3.2. Then in Section 4.3.3 we will present an evaluation of the approaches.

4.3.1 Experimental Dataset

Clinical records are often in short supply, with various ethical implications to consider. Tasks such as the i2b2 Challenge provide a rare chance to work with such annotated data. However, in the early stages of this work the i2b2 dataset was unavailable, therefore other avenues were sought for data. BLULab facilitates clinical NLP research through its NLP repository, subject to Data Usage Agreement. BLULab provided a subset of discharge summaries and progress reports to the i2b2 challenge. For our experimentation, we annotated a subset of History and Physical (H&P) examination reports. H&P reports were chosen by the authors as they were found to be more expressive and aim to created a more complete patient profile, e.g. family, work and personal circumstances. The primary aim of this dataset was to evaluate our approaches to Temporal Grounding and Condition Attribution.

4.3.1.1 Annotating Medical Conditions

For the tasks of Condition Attribution and Temporal Grounding, 120 H&P reports were randomly extracted from the BLULab’s NLP repository. As already

\[\text{http://www.dbmi.pitt.edu/nlpfront}\]
stated, this report type’s fuller descriptions make it richer than previous datasets in instances of condition attribution and temporal grounding. A breakdown in the distributions of these annotations can be seen in Tables 4.1 and 4.2.

H&P reports may vary in the organisation of content, but the content is mostly uniform, expressing the same information about patients (Sager et al., 1987). As well as this, many reports feature headings for different sections of the report (past medical history, impression), information which can be used as features in a classification task. Before annotating conditions found in the text, preprocessing was required in order to retain such information.

<table>
<thead>
<tr>
<th>Class</th>
<th>Count</th>
<th>Class</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>872</td>
<td>Historical</td>
<td>448</td>
</tr>
<tr>
<td>Other</td>
<td>93</td>
<td>Recent</td>
<td>424</td>
</tr>
<tr>
<td>Total</td>
<td>965</td>
<td>Total</td>
<td>872</td>
</tr>
</tbody>
</table>

Table 4.1: Condition Attribution Instances in the UCD Condition Attribution and Temporal Grounding Corpus

Table 4.2: Temporal Grounding Instances in the UCD Condition Attribution and Temporal Grounding Corpus

Preprocessing of the data consisted of a Java program that extended LingPipe\(^2\) tools in order to correctly split sentences on this dataset, and extract the heading for the section in which the sentence is contained.

The preprocessing outputs the sentence number, followed by a separator, the sentence’s contents and the heading under which the sentence features. Sentences were split for ease of annotation and also to allow parsing and part-of-speech tagging by the C&C\(^3\) parsing tools. C&C was chosen for its scalability, speed and the accuracy of its biomedical language models. As C&C does not provide a sentence splitter, LingPipe’s\(^4\) splitter was availed of for this task.

Annotation of the dataset was performed by two annotators over a 60 hour period. The annotator was presented with the collection, to annotate with an XML like tag “CONDITION”. This tag must have two attributes, “EXP” representing condition attribution and “HIST” representing the temporal grounding of the condition.

- **HIST**: A value of 1 indicates the occurrence of a historical condition, where 0 describes a current or recent condition. e.g. “The patient presen-

\(^2\)http://alias-i.com/lingpipe/

\(^3\)http://svn.ask.it.usyd.edu.au/trac/candc

\(^4\)http://alias-i.com/lingpipe/
ted with <CONDITION NUM="1" HIST="0">renal failure</CONDITION>” would indicate the condition “renal failure” is current.

• EXP: A value of 1 implies the experiencer is the patient with 0 signifying “other”, e.g., “The patient has a family history of <CONDITION NUM="1" EXP="0">hypertension</CONDITION>” signifies the condition “hypertension” is not experienced by the patient.

4.3.2 System Descriptions

Prior to feature extraction some pre-processing was required. The first stage of the process required sentences to be split. This was achieved through extending LingPipe tools. Initial experimentation also included the Genia sentence splitter\(^5\) however through cursory analysis we identified LingPipe as the most suitable tool. The second stage of preprocessing performed POS-tagging and dependency parsing using the C&C tools\(^6\). C&C tools were chosen for their scalability and biomedical language models allowing for simpler porting to future, larger datasets.

The aim of this approach is twofold. First, to identify an optimal feature-set for the task of assertion classification. Secondly, to identify a machine learning algorithm best suited to the task. In the following subsections we present the machine learning algorithms investigated, as well as the feature sets explored.

4.3.2.1 Machine Learning Algorithms

Early work in the resolution of assertion status primarily focused on the use of manually created rule-based systems, with more recent work focusing on statistical and ML methods. However, the domain of ML contains many sub-paradigms and varying approaches to classification. In this set of experiments, four ML methods that have not been previously applied to this task are explored. These four classifiers, namely Naive Bayes, k-Nearest Neighbour, Decision Trees and Random Forest represent the paradigms of probabilistic, lazy, statistical and ensemble learning, respectively. To provide a comparison to approaches previously used in the domain, the SVM is also evaluated.

Naive Bayes is a probabilistic classifier implementing Bayes Theorem. As a result, features implemented using this classifier are deemed to be independent.

\(^5\)http://www.nactem.ac.uk/y-matsu/geniass/
\(^6\)http://svn.ask.it.usyd.edu.au/trac/candc/
Despite this strong assumption it has been shown to be more than successful in text classification tasks such as spam filtering (Provost, 1999).

k-Nearest Neighbour (kNN) (Cover and Hart, 1967) is a simple pattern recognition algorithm that classifies an instance according to its distance to the \( k \) closest training instances. This algorithm has been chosen to represent the paradigm of lazy learning, i.e. there is no training phase as all computation is performed at the classification stage. Despite its simplicity, k-NN has often produce high accuracy results in comparison to other approaches (Caruana, 2006).

Decision Trees are a predictive model where the label of an instance is predicated on the system making decisions on certain feature values, determined by their Information Gain (Quinlan, 1993). For our experimentation, we utilise the C4.5 algorithms owing to its allowance of continuous attributes.

The key idea of the SVM is to identify a hyperplane that separates vectors in the data set into two classes such that the distance between the nearest point from each class to the hyperplane, is a maximum, known as the maximum margin. The SVM represents the state of the art in binary classification tasks.

The final classifier chosen for this task represents the state-of-the-art in machine learning, namely the Random Forest algorithm (Breiman, 2001). A Random Forest consists of many different decision trees, combining bagging (Breiman, 1996), and random feature selection.

4.3.2.2 Features

In this section, a list of features contributing to this task are presented. All features are automatically extracted using a set of tools developed by the authors.

4.3.2.3 Trigger List Features

Trigger list features are binary checks to determine if a keyword occurs within a certain span of the condition being inspected. The features implemented are presented in Table 4.3.
<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>precededByHistTerm</td>
<td>This feature performs a look-up for trigger terms from the historical word list, checking if it directly precedes the condition. An example historical trigger term would be “history of” as in “a history of diabetes”. If a condition, such as diabetes, is modified by a historical trigger term, it will return 1, otherwise 0.</td>
</tr>
<tr>
<td>containsHistMention</td>
<td>This is a weaker form of precededByHistTerm, checking simply if a trigger term from the historical list occurs in the same sentence as the condition. If one does, it will return 1 otherwise 0.</td>
</tr>
<tr>
<td>hasRelativeMention</td>
<td>If the sentence which contains the condition also contains a trigger term from the experiencer list such as ‘mother’, ‘father’ or ‘uncle’ it will return 1, otherwise 0.</td>
</tr>
<tr>
<td>hasPatientMention</td>
<td>1 if the sentence mentions the patient, otherwise 0.</td>
</tr>
<tr>
<td>containsDeath</td>
<td>1 if it contains the terms “deceased”, “died” from the death trigger terms list otherwise 0. A sentence describing death is more likely to refer to a relative, rather than the patient.</td>
</tr>
<tr>
<td>mentionsCommunity</td>
<td>1 if one of “area”, “community” from the geographical trigger list is mentioned, otherwise 0. If a sentence describes a community, as in “there has been a recent outbreak of flu in the area”, it is not referring to the patient, therefore the condition should not be attributed to the patient.</td>
</tr>
<tr>
<td>precededByWith</td>
<td>1 if the condition is directly preceded by “with”, otherwise 0. “with” was found to have higher frequency when describing patients rather than individuals other than the patient. e.g. “Patient presented with high blood pressure and fever.”</td>
</tr>
<tr>
<td>containsPseudoTerms</td>
<td>Pseudo-historical terms or phrases may mention a term that is found in the Historical list, but do not indicate that a condition mention in the same sentence is being used in a historical context. For example, “poor history” is a pseudo-historical trigger term. It uses a historical trigger term (“history”); however “poor history” refers to the incomplete nature of the patient’s medical history, not the historical nature of their condition. This feature returns 1 on the occurrence of a pseudo trigger term, otherwise 0.</td>
</tr>
</tbody>
</table>

Table 4.3: Trigger List Feature Descriptions for Binary Assertion Classification

### 4.3.2.4 Score based features

Scored based features used in this system extend and reinforce Trigger List features by computing a normalised score for the number of occurrences of Trigger List terms\(^7\). This feature aims to add further granularity to the de-

---

\(^7\)These trigger lists may be downloaded at http://csserver.ucd.ie/~jcogley/downloads/wordlists.tar.gz
cision making of the machine learning algorithms, presenting a floating point number rather than a binary feature.

The equation for computing these scores is defined as follows.

\[
s = \frac{N_t}{(N_w - S_w)}
\]  

(4.1)

\(N_t\) represents the number of trigger terms found in the sentence that contains the condition, \(N_w\) is the total number of words in the sentence, with \(S_w\) being the number of stopwords. These scores are calculated for each type of trigger term. For example, for trigger type relative_mention, a score is calculated using mentions of relatives in the sentence.

### 4.3.2.5 Contextual features

In resolving the textual context of conditions, it is important to look at what surrounds the condition beyond the lexical items. With these contextual features, we can capture that section in which a sentence occurs, and how many conditions occur in the sentence. Feature descriptions are presented in Table 4.4.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>isInFamHist</td>
<td>The importance of header information is motivated by the assumption that conditions that fall under explicit headings, are more than likely to have a greater affinity to the heading. This feature returns 1 if it is under Family History, 0 otherwise.</td>
</tr>
<tr>
<td>isInList</td>
<td>A binary feature denoting whether a condition occurs as part of a list of conditions, with one condition per line. Returns 1 if it is a member of such a list, otherwise 0.</td>
</tr>
<tr>
<td>numOfConditions</td>
<td>This feature represents the number of conditions present in a given sentence. A higher number of conditions indicates that the condition may be part of a list. Sentences that contain a list of conditions tend to list past conditions rather than recently suffered illnesses.</td>
</tr>
</tbody>
</table>

Table 4.4: Contextual Feature Descriptions for Binary Assertion Classification

### 4.3.2.6 Linguistically motivated features

These features were designed to monitor the effect of the verb tense on a

---

8The list of stopwords may be downloaded at [http://csserver.ucd.ie/~jcogley/downloads/stopwords.txt](http://csserver.ucd.ie/~jcogley/downloads/stopwords.txt)
condition. This feature has already been shown to aid the classification process (Mowery et al., 2009). For this task, linguistic features were extracted from the output of the C&C parsing tool, using both part-of-speech tags along with dependency information. The features are presented in Table 4.5.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>hasPastTense</td>
<td>A binary feature with 1 indicating the sentence contains a past tense verb, 0 otherwise. e.g. “The patient previously suffered renal failure” would return 1. If a condition is modified by a past tense verb, it has occurred in the past.</td>
</tr>
<tr>
<td>hasPresentTense</td>
<td>A binary feature with 1 indicating the sentence contains a present tense verb, 0 otherwise. If a condition is modified by a present tense verb, the condition is current. e.g. “the patient presents coughing”.</td>
</tr>
<tr>
<td>containsModalVerb</td>
<td>A binary feature with 1 indicating the sentence contains a modal verb, 0 otherwise. e.g. “palpitations may have been caused by anxiety”. The presence of the modal “may” following the condition indicates the condition is currently being examined and is therefore recent.</td>
</tr>
<tr>
<td>tenseInClause</td>
<td>Analyses the tense found in the same syntactic clause as the condition being examined. For example, in “abdominal pain has ceased, but patient now complains of lower extremity pain”, “abdominal pain” has a past tense within its clausal boundary, where the clause which contains “lower extremity pain” has a present tense verb.</td>
</tr>
<tr>
<td>tenseChange</td>
<td>Determines whether the verb tense used in the clause that contains the condition differs with the verb in another clause in the sentence. e.g. “Migraines persist yet palpitations resolved”. This feature allows finer granularity in resolving the tense surrounding conditions, such as the description of current conditions in the context of the patient’s history.</td>
</tr>
</tbody>
</table>

Table 4.5: Syntactic Feature for Binary Assertion Classification

4.3.3 Evaluation Results

In this section, we present the performance results of the Temporal Grounding and Condition Attribution systems described in Section 4.3.2, with respect to the aforementioned evaluation metrics on the UCD Temporal Grounding & Condition Attribution Corpus. Both machine learning systems are evaluated in comparison to the rule-based ConText system (Chapman et al., 2007). ConText is a rule-based system that uses pattern matching and trigger terms to determine the assertion status of a medical condition. This system has been
shown to be more than capable in negation detection and will be treated as a baseline for Temporal Grounding and Condition Attribution tasks.

The experiments have three key purposes. Firstly, to compare the performance of machine learning and manual rule-based methods. Both methods have their advantages, manual systems are usually high-performing given a limited domain. However, expanding and adapting a rule-based system is a costly endeavour, both in terms of time and effort. Given their adaptability, machine learning approaches are often preferred. Constant development of machine learning methods has seen near-perfect performance across several application domains. However, one issue within machine learning methods is the requirement of training data in the application domain. Given the focus of this work on the clinical domain, this is a more pertinent issue given data availability restrictions as discussed in Chapter 2.

Secondly, a multitude of machine learning algorithms are available for classification tasks. In these experiments a range of supervised machine-learning algorithms are investigated as described in Section 4.3.2.1.

Furthermore, the experiments provide the opportunity to analyse the contribution of each feature set to the classification tasks. While approaches may incorporate millions of features to achieve state-of-the-art performance, they are highly computationally expensive. By identifying the most effective feature-sets, we may focus efforts on those feature-sets, reducing the overall feature-set size and therefore the need for expensive computations.

To explore the optimal feature-set, all permutations of feature types as described in Section 4.3.2.2. The feature combinations and run identifiers are outlined in Table 4.6.

In the next subsections, we present the evaluation of Condition Attribution and Temporal Grounding performance.

### 4.3.3.1 Condition Attribution Experiments

Condition Attribution is a binary classification problem. We are interested in determining if a medical condition affects the patient described in a medical report, or some other individual, such as a family member. The dataset used for this experimentation is as described in Section 4.3.1. As a result of this dataset’s relatively small size of \(\approx 1000\) instances, we perform two stages of experimentation. The first set of experiments use 10-fold cross validation in
Table 4.6: UCD Temporal Grounding & Condition Attribution Feature-Set Combinations

<table>
<thead>
<tr>
<th>ID</th>
<th>Feature-Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>TrigLingConScore</td>
<td>trigger, linguistic, score-based, contextual</td>
</tr>
<tr>
<td>TrigLingScore</td>
<td>trigger, linguistic, score-based</td>
</tr>
<tr>
<td>TrigLingCon</td>
<td>trigger, linguistic, contextual</td>
</tr>
<tr>
<td>TrigConScore</td>
<td>trigger, score-based, contextual</td>
</tr>
<tr>
<td>LingConScore</td>
<td>linguistic, score-based, contextual</td>
</tr>
<tr>
<td>TrigLing</td>
<td>trigger, linguistic</td>
</tr>
<tr>
<td>TrigScore</td>
<td>trigger, score-based</td>
</tr>
<tr>
<td>TrigCon</td>
<td>trigger, contextual</td>
</tr>
<tr>
<td>LingScore</td>
<td>linguistic, score-based</td>
</tr>
<tr>
<td>LingCon</td>
<td>linguistic, contextual</td>
</tr>
<tr>
<td>ConScore</td>
<td>score-based, contextual</td>
</tr>
<tr>
<td>Trigger</td>
<td>trigger</td>
</tr>
<tr>
<td>Ling</td>
<td>linguistic</td>
</tr>
<tr>
<td>Score</td>
<td>score-based</td>
</tr>
<tr>
<td>Con</td>
<td>contextual</td>
</tr>
</tbody>
</table>

Table 4.7: Condition Attribution: Cross-Validation Results of Classification Algorithms on UCD Dataset

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>K-Nearest Neighbour</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>SVM</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Majority Class</td>
<td>0.9</td>
<td>1.0</td>
<td>0.95</td>
</tr>
</tbody>
</table>

order to understand the generalisability of the approaches. Further to this, the dataset was split into hold-out sets in order to determine how well the system can adapt to unseen data.

The dataset used to evaluate Condition Attribution is highly skewed in favour of conditions affecting the patient, where only 10% of conditions are experienced by another individual. As a result of this high skew, we see in Table 4.7 that a Majority Class classifier that classifies all instances as members of the majority class achieves an f-score of 0.95. With the high skew among the dataset, each algorithm could achieve perfect performance across the folds.

To gauge the difficulty of this task, we measure its performance against the ConText rule-based system. The evaluation was performed on a hold-out
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>K-Nearest Neighbour</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>SVM</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>ConText</td>
<td>1.0</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>Majority Class</td>
<td>0.9</td>
<td>1.0</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Table 4.8: Condition Attribution: Performance Results on 30% Training Data Vs. ConText on UCD Dataset

![Graph showing feature-set precision performance for condition attribution](image)

Figure 4.2: Feature-Set Precision Performance for Condition Attribution

set, through splitting the annotated dataset. The machine learning algorithms were trained on 30% of the dataset, with all systems evaluated on the remaining 70%. The results presented in Table 4.8 show that the ConText approach found the task also relatively simple. All machine learning approaches achieve perfect performance, with ConText misclassifying one instance.

While the approaches above may achieve perfect performance, one may posit that the feature-set is not optimal in that there may be features that contribute nothing to the task. To analyse this, we permute the feature sets as shown in Table 4.6. The results are presented in Figure 4.2.

In permuting across the feature sets, we may ascertain the contribution of each feature, as well as identifying those which combine well. A common trend among the feature sets is that the inclusion of Contextual features, Con, sees each algorithm achieve perfect performance, including the runs that contain only contextual features. The strength of this feature set lies with the feature isInFamHist which checks if the instance occurs under the Family History
The contribution of this feature indicates that occurrences of a condition affecting another individual, described outside of this section are rare.

The performances of each of the machine learning algorithms are largely similar across feature sets. As has been shown among the machine learning literature, Random Forests exhibit state-of-the-art performance that is robust to poor performing features, through selection of the best performing features for a given task. An exception to this is the combination of Linguistic and Score features, which favoured the use of the decision tree. This is in contrast to the decision tree’s relatively poor performance in using just score based features, an expected conclusion in the use of decision trees with floating-point values. Although the C4.5 algorithm may support floating-point features, it does not account for them adequately. Despite small differences in performance between feature sets, the contribution of Contextual features must be highlighted, so much so that rule-based systems may suffice for this binary classification task.

4.3.3.2 Temporal Grounding Experiments

Temporal Grounding is treated as a binary classification problem, similar to Condition Attribution. There are two states, one where the condition is currently affecting the patient, the other where the patient was admitted with the condition previously, such that it no longer affects the health of the patient. The same approach is used in experimentation as described in Section 4.3.3.1. First, we analyse the performance of several machine-learning algorithms using cross validation, following this we evaluate performance on a hold-out dataset in comparison to the ConText rule-based system.

The split between positive and negative instances were much more evenly distributed for the Temporal Grounding task than the Condition Attribution task. As a result, the f-score for the Majority Class classifier attained only 0.39. The 10-fold cross-validation results are shown in Table 4.9. The state of the art Random Forest attained the highest f-score, outperforming the nearest performing system, the Decision Tree by a statistical significant margin. While the Random Forest achieved the highest recall and f-score, its precision was outperformed by the Naive Bayes classifier. Similarly, the SVM achieved comparable recall. However, both the SVM and Naive Bayes were severely hampered by their poor precision and recall performance, respectively.

\*Significance is calculated using the Paired Student’s t-test with a p-value = 0.05, for all experiments in this chapter
Table 4.9: Temporal Grounding Cross-Validation Results of Classification Algorithms on UCD Dataset. † indicates the best system by a statistically significant margin. ‡ indicates a statistically significant difference in performance than the next ranked system.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest†‡</td>
<td>0.85</td>
<td>0.86</td>
<td>0.85</td>
</tr>
<tr>
<td>Decision Tree‡</td>
<td>0.85</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>K-Nearest Neighbour</td>
<td>0.83</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>SVM‡</td>
<td>0.76</td>
<td>0.86</td>
<td>0.80</td>
</tr>
<tr>
<td>Naive Bayes‡</td>
<td>0.91</td>
<td>0.58</td>
<td>0.71</td>
</tr>
<tr>
<td>Majority Class</td>
<td>0.30</td>
<td>0.55</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Table 4.10: Temporal Grounding Performance Results on 30% Training Data Vs. ConText on UCD Dataset. † indicates the best system by a statistically significant margin. ‡ indicates a statistically significant difference in performance than the next ranked system.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest†‡</td>
<td>0.82</td>
<td>0.86</td>
<td>0.84</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.81</td>
<td>0.84</td>
<td>0.82</td>
</tr>
<tr>
<td>K-Nearest Neighbour‡</td>
<td>0.80</td>
<td>0.82</td>
<td>0.81</td>
</tr>
<tr>
<td>SVM‡</td>
<td>0.73</td>
<td>0.86</td>
<td>0.79</td>
</tr>
<tr>
<td>Naive Bayes‡</td>
<td>0.92</td>
<td>0.61</td>
<td>0.74</td>
</tr>
<tr>
<td>ConText</td>
<td>0.8</td>
<td>0.61</td>
<td>0.69</td>
</tr>
<tr>
<td>Majority Class</td>
<td>0.55</td>
<td>1.0</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Following the Cross-Validation experiments, each algorithm is evaluated on a hold-out set as in Section 4.3.3.1. Through ranking the systems by f-score in Table 4.10, the performance of systems mirror that of the cross-validation experiments. Again, Naive Bayes achieved the greatest precision but was severely hampered by its recall. The SVM matched the Random Forest’s recall, but had poor precision. All machine learning approaches outperformed ConText, which had competitive precision but was let down by its recall. This highlights the problem of poor adaptability among manual rule-based approaches.

Analysing the permutations of feature sets allows us to determine an optimal feature set for the task. Figure 4.3 shows precision performance of each algorithm using feature set combinations on the hold out set as detailed above. The first thing to stand out is the high precision performance of Naive Bayes over all other investigated algorithms. All algorithms bear the same pattern, even if performance between algorithms differs. Firstly, any feature set containing score-based features achieved top performance. As stated in Section 4.3.2.2, score-based features calculate a normalised score for trigger terms in the text,
the result being a score indicating the impact or weight of that trigger term in the sentence. Rather than using trigger term features that were binary encoded, score based features offer more granularity, highlighted by precision improvements. Moreover, trigger features are the least precise, with runs in which they contribute representing the worst performing runs in terms of precision.

In Figure 4.4, we detail the recall performance across feature-sets. Again, Naive Bayes stands out, however on this occasion due to its poor recall. The graph highlights the precise nature of score based features, with runs consisting of score based features having the poorest recall, especially for the SVM and Naive Bayes approaches. Despite these approaches posting lower scores, the Decision Tree coupled with score based features achieved highly competitive performance. We see also the power of the more general trigger based features. The most positive recall performance across all algorithms was Trig which uses only trigger features. This is consistent throughout the feature combinations, with the inclusion of trigger features correlating with top recall performance in runs.

In Figure 4.5, we see the full contribution of the feature sets and algorithms in the shape of their f-score. The poor recall of Naive Bayes damages its performance irreparably. In a similar fashion, the total performance of the SVM is severely hampered by its poor precision, notably in its use of score based features. The Random Forest was by far the most consistent approach, outperforming or matching all other algorithms across feature sets, highlighting its robustness through feature selection. The often overlooked lazy method,
Figure 4.4: Feature-Set Recall Performance for Temporal Grounding

Figure 4.5: Feature-Set F-score Performance for Temporal Grounding
k-nearest neighbour, has much to contribute to the task with the identification of an optimal feature-set keeping computation tractable.

**What features contribute the most?**

Across the f-scores, the score based measures contribute the most to the task. While the trigger features did not post the most impressive f-score, their importance to the recall of systems must not be overlooked. Linguistic features, which contain information about linguistic content such as tense are also important to the task. For example, its contributions to runs containing trigger and score based features improve performance from competitive to top performance. While contextual features contribute much to the task of Condition Attribution, this is not the case for Temporal Grounding. Though historical information may typically be found in a history section of a medical record, it may well be littered across the text in order to provide context to a patient’s current condition. Its weak contribution is perhaps best highlighted through its removal from the run TrigLingConScore, making no impact on performance of the run TrigLingScore.

### 4.4 Assertion Classification on the i2b2/VA dataset

In this section, we present on Assertion Classification system that builds on the tasks of Condition Attribution and Temporal Grounding to create a multi-class classification system, evaluated on the 2010 i2b2/VA challenge corpus. In Section 4.4.1, we present an overview of the i2b2/VA assertion classification corpus. Section 4.4.2 presents a description of the system. In Section 4.4.3, we provide details of an evaluation of the system.

#### 4.4.1 Dataset Overview

The dataset for the multi-class Assertion Classification task is from the same report base as that of the Concept Identification task. Only the annotations differ. The task of Assertion Classification relates to problem concepts. Clinical narratives contain many mentions of medical problems however not all directly affect the patient who is the focus of the narrative. We note that the Assertion Status of a medical problem denotes its status of affecting the patient. The i2b2 task Assertion Classification task allows six possible assertion
statuses: Present, Absent, Possible, Conditional, Hypothetical and Not Associated with the Patient (NotPatient). Examples of these assertion status are provided in Table 4.11.

<table>
<thead>
<tr>
<th>Status</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>This morning she developed fatigue.</td>
</tr>
<tr>
<td>Absent</td>
<td>She denies any cough / fever.</td>
</tr>
<tr>
<td>Possible</td>
<td>He also had a high CO and low SVR suggestive of sepsis.</td>
</tr>
<tr>
<td>Conditional</td>
<td>Pain with movements of his back .</td>
</tr>
<tr>
<td>Hypothetical</td>
<td>Patient may experience vomiting, fevers</td>
</tr>
<tr>
<td>NotPatient</td>
<td>Father suffered from Type II Diabetes</td>
</tr>
</tbody>
</table>

Table 4.11: Examples of Assertion Status in the i2b2 Challenge Corpus

The Assertion Status of medical problems were annotated by human annotators as part of the dataset provided by the i2b2 challenge. A breakdown of the annotations provided for the Assertion Classification task are detailed in Table 4.12.

<table>
<thead>
<tr>
<th>Status</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>4,624</td>
</tr>
<tr>
<td>Absent</td>
<td>1,596</td>
</tr>
<tr>
<td>Possible</td>
<td>309</td>
</tr>
<tr>
<td>Conditional</td>
<td>73</td>
</tr>
<tr>
<td>Hypothetical</td>
<td>382</td>
</tr>
<tr>
<td>NotPatient</td>
<td>89</td>
</tr>
</tbody>
</table>

Table 4.12: Assertion Status Counts in i2b2 Challenge Corpus

4.4.2 System Description

The previously discussed Condition Attribution and Temporal Grounding systems perform binary classification, however the task as defined for the i2b2/VA challenge is a multi-class problem. Multi-class classification aims to classify each instance into one of N classes. This differs from more traditional classification tasks, where there are two possible classes as in Section 4.3.3. To gauge the difficulty in performing multi-class classification, we first apply feature sets from the Temporal Grounding and Condition Attribution system with the top performing algorithm, Random Forest, as shown in Table 4.13.

Notably performance is greatest for the Present, Absent and Hypothetical classes. These three classes represent over 90% of the dataset and as such, have an overwhelming bias in classification performance. For the further classes performance is extremely poor. Conditional and NotPatient instances represent a little over 1.5% of the dataset. Heightened precision for non-patient instances can be attributed to the isInFamHist feature. Instances of class Possible are as numerous as the Hypothetical class, however instances of
<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>0.83</td>
<td>0.95</td>
<td>0.89</td>
</tr>
<tr>
<td>Absent</td>
<td>0.86</td>
<td>0.71</td>
<td>0.78</td>
</tr>
<tr>
<td>Hypothetical</td>
<td>0.82</td>
<td>0.79</td>
<td>0.80</td>
</tr>
<tr>
<td>Conditional</td>
<td>0.13</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>Possible</td>
<td>0.43</td>
<td>0.11</td>
<td>0.17</td>
</tr>
<tr>
<td>NotPatient</td>
<td>0.89</td>
<td>0.37</td>
<td>0.52</td>
</tr>
<tr>
<td>Micro Avg.</td>
<td>0.83</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>Macro Avg.</td>
<td>0.66</td>
<td>0.49</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Table 4.13: Results of extending Binary feature-set to Multi-Class Assertion Classification Using Random Forest

- Depression Vs. Borderline personality disorder.
- The idea of a perineoplastic syndrome has reared its head.

Figure 4.6: Examples of the class Possible in the i2b2 Assertion Classification task

Possible are quite nuanced in their expression as in Figure 4.6. Due to the high similarity of Present and Possible instances, as well as the level of data skew, it is a common error for possible instances to be classed as Present.

To this point, no feature specifically targets these new classes, hence its poor performance. To account for these new classes, and possibly more for future directions, we incorporate specific pattern based features as well as more generalisable token sequence features. The first extension made to the system is the introduction of new mention and score features. As in Section 4.3.2.2, trigger lists are used. The system on the i2b2/VA dataset extends these trigger lists to account for the additional assertion classes. New features were developed as described in Table 4.14.

Furthermore, we incorporate token features to create generalisable patterns among the texts. With the development of these new features, the feature-set comprises of $\approx 300,000$ features owing mainly to the inclusion of tokens as features. Initial attempts at experimentation using this feature set created huge runtime issues for Weka. Further to this, leading systems have been shown to implement highly optimised Support Vector Machines. To this end, we use the LIBSVM package\(^\text{10}\) to perform our training and classification due to its high-performance and training efficiency.

\(^{10}\)http://www.csie.ntu.edu.tw/~cjlin/libsvm/
<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>containsAllerg</td>
<td>A binary feature with 1 indicating the problem contains the string “allerg” as a prefix, 0 otherwise. Allergies are a common Conditional problem, with issues arising due to an allergic reaction.</td>
</tr>
<tr>
<td>someoneElseAgePattern</td>
<td>The pattern &quot;COND at AGE&quot; is typical of conditions that are experienced by other individuals. This binary feature identifies if the condition matches this pattern.</td>
</tr>
<tr>
<td>surroundedByQuestionMark</td>
<td>Possible conditions are often preceded/succeeded by “?”. This binary feature indicates the presence of the question mark before or after the inspected condition.</td>
</tr>
<tr>
<td>surroundedByComma</td>
<td>Conditions found in lists are typically negated. To leverage this information, this binary feature indicates if a medical condition is preceded and followed by a comma.</td>
</tr>
<tr>
<td>containsConditionalPattern</td>
<td>Conditional conditions are frequently followed by tokens such as “with”, “upon”, “when”, “on” while also preceded by a set of conditional phrases. This feature indicates if the condition fits the conditional pattern.</td>
</tr>
</tbody>
</table>

Table 4.14: Features added to Multi-class Assertion Classification approach

### 4.4.3 Evaluation Results

In this section, we present the results of our experiments in which we extend the binary classification Temporal Grounding and Condition Attribution systems to the multi-class assertion classification task at the Fourth i2b2/VA challenge.

**How do we extend binary classification to multi-class classification?**

As stated in the previous section, the Assertion Classification system uses an SVM. A problem for the SVM however, is the requirement for multi-class classification. To apply a binary classifier to a multi-class problem, we must partition the problem into a subsets of binary problems, then combine the results of the systems to achieve our multi-class classification goal. There are two schools of thought in performing this task, One Versus All (OVA) or One Versus One (OVO). OVA classification creates a binary classifier for each class. In each of these classifiers an instance may be positive (is an instance of the class) or negative (is an instance of any other class). For OVO classification, a binary classifier is created for each possible pair of classes e.g. Present Vs. Absent. For both approaches, the outputs are combined through picking the highest confidence score among each of the classifiers, in the case of the SVM its probability estimates are used. The literature discussion on the topic of which ap-
The use of the SVM with the new feature set has shown a marked improvement over the previous approach using the Weka toolkit, in particular on the minority classes. On the note of OVO Vs. OVA, there are some points of interest. OVO favours higher recall, with its recall besting OVA in three of the classes. However, this is to the detriment of its precision in some cases. Despite this, all of our further experiments will focus on the usage of the OVO approach as we wish to increase our recall on minority classes. Issues with precision of the system may be rectified through kernel choice and optimisation of the SVM.

What is the correct kernel?

In Table 4.16 we evaluate kernel choice for our problem domain. The RBF kernel shows precision gains across the classes, save Present, where it in fact boosts its recall over that of the linear kernel. Overall the RBF kernel posts a higher f-score in all but two classes, increasing both micro and macro averages. This overall performance gain, though minimal in some cases, leads our experiments to continue with the usage of an RBF kernel. The results outlined in Table 4.16 were achieved with an optimised cost parameter of 100,000 and gamma set to 0.01.
What features contribute to the task?

As well as optimising the algorithm, we aim to discover the impact of certain feature groupings in multi-class assertion classification. In Table 4.17 we see the micro and macro averages of scores in assertion classification. As we continue to add feature sets we see performance continuing to increase. Our final feature set +context, which combines all feature groupings shows a 9% increase over the baseline token feature set, and a 20% increase in Macro-Average f-score over the initial feature-set described in Table 4.13.

Though we see an overall gain in performance, we must further analyse the effects of each feature set on individual classes, as this was a failing of the Weka system in particular on minority classes. In Table 4.18 we present a breakdown of feature-set performance for each class.

The first thing we learn is that surface based features, i.e. tokens contribute much to the task. Though treated as a baseline, token features using the optimised SVM thoroughly outperforms the initial approach utilising Weka. This indicates that patterns of tokens prior to, and after a concept are generalisable to a point to generate positive results. The introduction of mention features further reinforces the use of pattern based information, as shown by the improvements for classes Absent, Hypothetical and NotPatient. Mention features perform checks for lists of trigger terms, as well as capturing certain patterns. For example, “no” or “absent” in the case of Absent class, or more specific patterns for Hypothetical “in case of X take Y”.

Heading information is pertinent across several classes, namely Possible, Conditional and NotPatient instances which all saw a growth in precision and recall. The importance of this feature set is heightened due to the high frequency of instances occurring under certain headings, i.e. Conditional under Allergies, NotPatient under Family History. This reiterates our finding in the our Condition Attribution experimentation in Section 4.3.3.1 where cases of NotPatient typically occurs under a family history section. The inclusion of score based features measure the impact a trigger term has on a concept. The use of this feature positively impacted particularly on Possible and NotPatient instances. However, there was little impact for other assertion classes. Notably it improves the precision of both of these classes where instances occur in short sentences, with a trigger term of its class, e.g. “Mother has diabetes”.

The final feature group included was the output of the ConText rule-based
<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Micro</td>
<td>Macro</td>
<td>Micro</td>
</tr>
<tr>
<td>token</td>
<td>0.899</td>
<td>0.817</td>
<td>0.899</td>
</tr>
<tr>
<td>+mention‡</td>
<td>0.910</td>
<td>0.839</td>
<td>0.910</td>
</tr>
<tr>
<td>+heading‡</td>
<td>0.915</td>
<td>0.891</td>
<td>0.915</td>
</tr>
<tr>
<td>+score‡</td>
<td>0.917</td>
<td>0.891</td>
<td>0.917</td>
</tr>
<tr>
<td>+context</td>
<td>0.919</td>
<td>0.883</td>
<td>0.919</td>
</tr>
</tbody>
</table>

Table 4.17: Top performing feature sets for i2b2 Assertion Classification, Micro and Macro Average. † indicates the best system by a statistically significant margin. ‡ indicates a statistically significant difference in performance than the next ranked system.

<table>
<thead>
<tr>
<th></th>
<th>Present</th>
<th>Absent</th>
<th>Possible</th>
<th>Conditional</th>
<th>Hypothetical</th>
<th>NotPatient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>token</td>
<td>0.906</td>
<td>0.977</td>
<td>0.912</td>
<td>0.874</td>
<td>0.807</td>
<td>0.378</td>
</tr>
<tr>
<td>+mention</td>
<td>0.912</td>
<td>0.976</td>
<td>0.932</td>
<td>0.897</td>
<td>0.807</td>
<td>0.404</td>
</tr>
<tr>
<td>+heading</td>
<td>0.912</td>
<td>0.981</td>
<td>0.947</td>
<td>0.895</td>
<td>0.812</td>
<td>0.412</td>
</tr>
<tr>
<td>+score</td>
<td>0.913</td>
<td><strong>0.982</strong></td>
<td>0.947</td>
<td>0.896</td>
<td><strong>0.825</strong></td>
<td><strong>0.413</strong></td>
</tr>
<tr>
<td>+context</td>
<td><strong>0.916</strong></td>
<td>0.979</td>
<td><strong>0.949</strong></td>
<td><strong>0.910</strong></td>
<td>0.824</td>
<td>0.409</td>
</tr>
</tbody>
</table>

Table 4.18: Top performing feature sets for i2b2 Assertion Classification, class breakdown
system. The ConText system may determine that an instance is present, negated, historical, hypothetical and not associated with the patient. The results of applying this as a feature are somewhat conflicting. For example, in the worst case, its inclusion significantly worsens the performance of the system on the NotPatient class. Conversely, it significantly improved the performance on the Hypothetical class. In the case of the Hypothetical class, the use of ConText refines performance of the mention feature which had previously seen large gains. Its inclusion lessens the impact of the heading feature-set on NotPatient instances, which with the high frequency of NotPatient instances occurring under Family History headings is detrimental to performance.

In analysing the contribution of feature sets, we have discovered the importance of identifying the context of a concept in determining its assertion status. Token based features as well as those which identify keywords and patterns surrounding the concept outperformed feature sets used in the Condition Attribution and Temporal Grounding as described in Section 4.3.3. Heading information greatly aided performance on Conditional and NotPatient classes, due to their high frequency of occurrences under Allergies and Family History headings, respectively. Runs using score features achieved top performance in Possible and NotPatient classes, indicating the regularity of shorter sentences as well as higher frequencies of mention terms in these classes. The inclusion of ConText improves performance in recognising Absent and Hypothetical instances, however at the expense of NotPatient and Possible classifications.

What are the system errors?

As previously stated in Section 4.4.1, the dataset is highly skewed in favour of the Present assertion class. This high data skew is a consistent source of errors, with 75.8% of errors made up of false positives of Present. This is a key cause of low recall scores for minority class instances such as Conditional and NotPatient, that while they exhibit high precision, the skew is far too high for the instances to be correctly classified.

The most numerous errors classified as Present were members of Absent and Possible. In the case of Absent misclassifications, lists were a frequent cause of issues. In Figure 4.7 we see a representative examples of Absent concepts classified as Present. The key error here is that no negation term is within the span of the ConText span check, nor the mention feature check. As a result of these insufficient span checks, negation was undetected and therefore each concept is classed as Present.
• He denied visual changes, light-headedness of dizziness currently, neck stiffness, or pain in any other area.

Figure 4.7: Example of Absent misclassified as Present

Despite our application of new patterns and trigger terms to the feature set to account for Possible instances, they still remain subtle. In Figure 4.8, the concept term ATN is to be classified as Possible, due to it being “felt to be” secondary to ischemia. The phrase “felt to be” is key to its classification as Possible, however this is often found in the context of tests where findings are described, so was excluded from our pattern matching rules.

• He developed renal failure felt to be ATN secondary to ischemia of his kidneys

Figure 4.8: Example of Possible Misclassified as Present

NotPatient instances were often confused with Absent cases due to close proximity of negation indicated through the ConText feature. For example, “No history of breast cancer” is determined to be negated rather than NotPatient as a result of the clear negation, though in spite of the fact that it was described under Family History. On removal of the ConText feature, these types of error are reduced as previously described in Table 4.18.

Though instances of Present false negatives were far less numerous, there are generalisable instances. Firstly, instances of Hypothetical often occur in a statement with a drug name e.g. “in case of X take Y”. However, the classifier over-generalised this occurrence to cases such as prescriptions “Y for X”, e.g. “Demerol for pain”. Here the pain is in fact present in the individual, rather than the drug being prescribed for pain that may occur.

The most common example of a Present false negative is the misclassification of Absent. This typically occurs in the results through having a matching span that is too great, where negation is incorrectly applied to items in lists. In Figure 4.9, we see an instance of this occurrence.

In this case, the “no”, which should be attached to complications is overgeneralised and incorrectly attached to the concept elevated right atrial pressure. This creates an interesting point in that for recognising Present conditions it is too general, but in the case of attempting to attach negation further in a list such as
Tests showed no complications, elevated right atrial pressure and moderate pulmonary hypertension.

Figure 4.9: Example of Present misclassified as Absent

in Figure 4.7, the criterion is too specific. Negation resolution through ConText is performed through patterns, as well as through span checking. However, in this case to correctly attach the negation particle, a higher form of analysis, such as dependency analysis must be used. This is not to say generic rules help, as we found in our initial evaluation of linguistic feature types as described in Section 4.3.3, they do not suffice in classification performance. However, syntactic dependency rules may be leveraged in order to correctly attach negation particles to items in lists may see an improvement in performance.

4.5 Summary

In this chapter we presented experimental evaluations of approaches to the task of Assertion Classification. The evaluations were performed on a dataset annotated as part of this work as well as the i2b2/VA dataset. The experiments provide answers to the following research questions:

How do rule-based systems compare to machine learning based approaches?

Our initial experimentation in Section 4.3.3 analysed the performance of our machine learning based system with respect to the rule-based system ConText for the binary classification tasks of Condition Attribution and Temporal Grounding. Due to large data skew in the Condition Attribution corpus, there was little difference in performance between machine learning approaches and ConText. However one point of note is the prevalence of conditions not associated with the patient occurring under the Family History heading, leading to the reasoning that sophisticated rule-based and machine learning systems may overcomplicate the Condition Attribution task. In the Temporal Grounding task, machine learning systems significantly outperformed the ConText system. This provides an insight that the resolution of temporal information requires deeper information than lexical cues.

With a multitude of machine-learning algorithms available for classification tasks, what algorithm performs best?
In the Condition Attribution task, there was no significant gains in using any algorithm, owing to high data skew and the high frequency of conditions not associated with the patient occurring under Family History. For the task of Temporal Grounding, several algorithms were investigated from varying theoretical backgrounds. While simpler systems such as Naive Bayes and K-Nearest Neighbour achieved high performance, they were outperformed by the robustness of the Random Forest classifier. In extending the approach to the i2b2/VA task, high computational costs of Weka due to a very large set of features, rendered the system ineffective. The LibSVM system was adapted for the approach due to its efficiency and high performance in the literature. The resulting system achieved performance comparable to the state of the art in the area.

**How may binary classification be extended to a multi-class Assertion Classification task?**

The initial approaches to Assertion Classification presented in this chapter treated the problem as a binary classification task. In moving toward multi-class classification, several system adaptations were required. Firstly, the feature set was required to be expanded to account for new classes. Secondly, to account for the expansion in feature set size, the highly efficient and optimised LibSVM was used. Finally, we performed evaluation of both OVO and OVA multi-class classification approaches. While both approaches have their merits, experimental findings showed higher recall on minority classes when using an OVO approach.

**What feature sets contribute most to the task of Assertion Classification?**

Feature analysis showed *heading* features to be the most important feature in the Condition Attribution task. In our analysis of the Temporal Grounding task, *our novel score-based features achieved near-optimal performance*, with the inclusion of trigger and linguistic features achieving optimal performance. In extending the approach to the multi-class problem, *token based features* formed the basis of the classification. The inclusion of mention, heading, score and information from ConText saw improvements in performance across the board. Given the nature of multi-class classification, there are several trade offs in applying certain feature sets. For example, the use of ConText as a feature saw a drop in performance in the *NotPatient* class, while aiding the recall of *Absent* classes.

In this chapter we analysed the performance of an Assertion Classification sys-
tem in the context of a shared task. This will lead us to apply this system as part of the Clinical Information Retrieval framework described in Chapter 6.
CHAPTER FIVE

INFORMATION RETRIEVAL IN THE MEDICAL WORLD

5.1 Introduction

In the preceding chapters of this thesis, the discussion and experimentation was focused on Natural Language Processing techniques in the Clinical domain. In this chapter and following chapters, the thesis will concentrate on Information Retrieval in the Clinical domain. Clinical data takes the form of free-text fields. Given these free-text datasets, traditional structured methods, e.g. relational databases, are not feasible, as they cannot account for unstructured text that contains varying and rich natural language. This leads to the use of Information Retrieval (IR) techniques that may retrieve documents based on a natural language query. In this chapter, we identify areas of interest in applying traditional IR models to the clinical domain. For example, the use of natural language in the clinical and biomedical domains causes issues as ambiguity is a common source of confusion in natural language texts, and clinical data that is rich in negation, temporal information and discussions relating to other individuals further complicates the task. To alleviate these issues, information extraction techniques may be used to expand on simpler keyword based search approaches. In this chapter, we provide an overview of Information Retrieval approaches in the biomedical and clinical domains while seeking answers to the following research questions:

- How are Information Retrieval systems evaluated?
- What specific problems arise in clinical and biomedical IR systems?
- What model is best suited for retrieving documents in the clinical and biomedical
domain?

• What strategies have been used in the literature to address domain specific issues?

• To what end have information extraction techniques been applied to IR in the biomedical and clinical domains?

In this chapter, we present an overview of Information Retrieval concepts that are integral in answering questions posed by this thesis. In Section 5.2, we present a discussion of the foundations of Information Retrieval, including systems facilitating the development of the Information Retrieval system used to answer our proposed research questions. Section 5.3 discusses Information Retrieval research in the biomedical domain. In Section 5.4, we provide an overview of research of Information Retrieval techniques, in the context of the clinical domain. In particular, we provide a survey of methods developed at the TREC 2011 and 2012 Medical Tracks to identify key problem areas in Clinical Information Retrieval.

5.2 Foundations of Information Retrieval

Information Retrieval (IR) is the task of satisfying an information need, expressed in a query, through the retrieval of a list of documents ordered by their relevancy to the information need of the user. A key difference between Information Retrieval and structured data retrieval methods, such as RDBMS, is the lack of structure. While IR may be performed across a number of unstructured media, e.g. text, sound, images, for the purposes of this thesis we focus on text-based retrieval.

In its most basic form, a query is a set of terms. Documents that contain these terms are then returned to the user. Searching across documents for term matches at query time is clearly an inefficient process. Therefore, documents are processed prior to querying to create an inverted index. Fundamentally, an inverted index maps each term found in the collection to the documents in which they occur. A simplistic, sample inverted index is shown in Table 5.1.

The simplest model of retrieval using such an index is the Boolean model. The Boolean model uses the operators AND, OR and NOT. For example, given the query “liver AND failure”, the documents doc1 and doc2 are retrieved.
PubMed\(^1\) is a prime example of the success of a Boolean Retrieval model, however simplistic. Issues with Boolean Retrieval include the fact that it does not provide a ranked list of documents ordered by their relevancy to an information need. Given any sizeable collection i.e., one that is not easily readable in its entirety by a user, this model is not an optimal solution in addressing a user’s information need. In the next section, we will describe models developed to rectify this issue.

### 5.2.1 Ranked Information Retrieval Models

In order to facilitate ranked retrieval models, inverted indexes may contain further information such as term frequencies, as in Table 5.2.

<table>
<thead>
<tr>
<th>Term</th>
<th>doc1</th>
<th>doc2</th>
<th>doc3</th>
</tr>
</thead>
<tbody>
<tr>
<td>liver</td>
<td>5</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>renal</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>failure</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>edema</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5.2: A Sample Inverted Index With Term Frequency Information

In Table 5.2 we extend the index in Table 5.1 to include the term counts for each document. For example, liver occurs five times in doc1 and four times in doc2, but has no occurrences in doc3. The collection of such statistics facilitates the calculation of measures of similarity between documents and queries, such that documents may be ranked according to these measures, as in the Vector-Space model.

#### 5.2.1.1 Vector-Space Model & tf-idf

In order to rank documents with respect to a query, a metric must be defined to calculate the similarity between a given query and a set of documents. The

---

\(^1\)http://www.ncbi.nlm.nih.gov/pubmed
term frequency-inverse document frequency (tf-idf) measure, an integral component of the vector space model, proposed and developed by Salton et al. (1975) is a seminal approach in ranked information retrieval. In this model, documents and queries are represented as term vectors. The tf-idf is calculated between these vectors as an indicator of the relevance of a document to a given query.

The tf-idf calculates a term weight using its frequency in a given document and its document collection frequency. To calculate the tf-idf score, we first count the term frequency denoted as \( tf(t, d) \), where \( t \) is the term and \( d \) is a given document. The second aspect of tf-idf is the inverse document frequency. The inverse document frequency is used as a measure of specificity (Sparck-Jones, 1972) of a given term and is calculated in Equation 5.1.

\[
idf_t = \log \frac{N}{df_t}
\]  

(5.1)

, where \( t \) is the term, \( N \) is total number of documents in a collection, \( df_t \) is the frequency of \( t \) in the given document.

The tf-idf of the term is the product of the \( tf \) and \( idf \) statistics. To calculate the similarity score between a query \( q \) and document \( d \), we sum the calculations as in Equation 5.2.

\[
sim(q, d) = \sum_{t \in q} tf-idf(t, d)
\]  

(5.2)

This continuous metric now allows for the ranking of the relevancy of documents. Though often an effective baseline measure, the tf-idf measure is unsuitable for the retrieval of clinical documents. Firstly, it performs poorly on large documents. Given the large variation of document lengths in clinical collections, as a lengthy visit will create more documentation as opposed to a check-up, this method is less suitable for a clinical retrieval approach. While there are methods to address document length issues (Singhal et al., 1996), the above methodology takes no account of term dependencies, so highly important information such as “colon cancer” may return the unrelated “ulcerative colon” as a term matches given that both spans contain the term “colon” but are otherwise unrelated.
5.2.1.2 Probabilistic Models

The Probability Ranking Principle (PRP) (van Rijsbergen, 1979) ranks documents by the probability that they satisfy a user’s information need. Original probabilistic methods utilised the Binary Independence Model (BIM) whereby term occurrences are marked through a binary method, as in boolean retrieval rather than term frequencies in the Vector Space model. The most commonly adapted probabilistic method is Okapi BM25 (Robertson et al., 1994), which leverages term frequencies, rather than the BIM approach. The Okapi model addresses issues of the Vector Space and BIM approaches through paying attention to the length of documents. Okapi BM25 is defined in Equation 5.3

\[
Okapi\_BM25(q,d) = \sum_{t \in q} idf(t,D) \times \frac{tf(t,d) \times (k_1 + 1)}{tf(t,d) + k_1 \times (1 - b + (b \times \frac{dl}{dl_{AVG}}))}
\] (5.3)

\(k_1\) is a tuning parameter that determines the role of term frequencies in ranking. \(k_1\) assumes the BIM, while a value of 1 indicates the usage of raw frequency counts. The \(b\) parameter determines the level of scaling by document length. Set to 0, no normalisation is performed, when set to 1 all term frequencies are normalised. \(dl\) and \(dl_{AVG}\) are the document length and average document length, respectively. Though we have improved on our suitability by considering this approach to address the problem of document length, the Okapi model takes no account of term dependencies.

5.2.1.3 Language Models

Language Models in IR are a further probabilistic school of thought in addressing the shortcomings of the Vector Space model representation (Ponte and Croft, 1998; Hiemstra, 2000). Language modelling approaches aim to rank each document by the probability that the document’s language model could output the query, \(P(Q, M_d)\), where \(Q\) is the query and \(M_d\) is the language model of a given document \(d\). Frequently referred to as \(n\)-gram language models, the probability of generated sequences of strings in a language is calculated through the previous \(n-1\) words observed. For \(n\) greater than one, we may account for term dependencies in the text.

At this point in the thesis, we have identified retrieval models that account for document length and term dependencies, which are key requirements in our
retrieval of clinical text. However, the described approaches do not facilitate
the use of a structured query language, which is essential when applying the
MER and Assertion Classification systems described in Chapters 3 and 4, re-
respectively. In the next section, we present our final investigated model, which
integrates language modelling and inference networks to allow for a struc-
tured query language while addressing the issues of document length and
term dependencies.

5.2.1.4 Language Models and Inference Networks

As stated in the previous section, language models may be utilised in account-
for term dependencies required to correctly identify a document as rele-
ant to an information need. Inference networks are based on Bayesian net-
works which are directed, acyclic graphical models. The first application of
Inference Networks to IR were presented by Turtle and Croft (1991) with the
method later adapted with language models (Metzler and Croft, 2004).

In the initial formulation of the inference network model, six node types are
available:

- **Document Nodes**: These nodes represent each document in the corpus. As
  part of this model, the standard representation is as multi-sets of binary
  vectors signifying the occurrence, or non-occurrence of a *feature*. A fea-
  ture may be any extractable concept from the text, allowing for the use
  of term dependencies or named entity information.

- **Smoothing parameter nodes**: These nodes accommodate the sparse data
  problem i.e., addressing the issue of sequences that have a probability of
  zero if the have not been seen before in the system.

- **Model nodes**: These nodes represent the model used in retrieving the doc-
  uments. The availability of different model nodes allow us to model dif-
  ferent sections of text in a given document.

- **Representation Nodes**: These nodes identify if a feature value defined in
  the Document node has occurred, e.g. “phrase liver failure occurred”.

- **Belief Nodes**: These nodes represent the structure of the query. They are
  used to combine *beliefs*. For example, the belief node may combine the
  representation nodes that indicate that the phrase liver failure occurred in
  a document along with the term acute.
• **Information Need Node**: This node combines all other information discovered in the network, which is then used to rank the documents.

Given its robust query language, its addressing of term dependencies, document length as well as other information generated from the previously discussed NLP systems, a language modelling and inference network approach will be adapted for our experimentation. In particular, we apply the Indri retrieval system (Strohman et al., 2005). A full description of the Indri system and its integration into our retrieval workflow will be discussed in Chapter 6.

### 5.2.2 Evaluation of Information Retrieval

Throughout this chapter we have described varying methods in the performance of Information Retrieval. The retrieval task requires the return of a ranked list of documents that are relevant to a user’s information need. In general, a top performing IR system is one such that it satisfies the user’s information need. However, such general terms are difficult to quantify. Therefore, the standardising of evaluation is a key aspect of IR research.

To evaluate Information Retrieval systems, there are three requirements: A test collection, a set of queries and a set of gold standard relevance judgements. Shared tasks such as TREC and NCTIR provide a forum for performing comparative evaluations as well as facilitating these three requirements. However, due to the large size of test collections as well as the domain knowledge required, especially in the medical domain, the development of test collections and relevance judgements are expensive. In this section, we will describe several metrics designed to quantifiably determine an optimal performing system while taking into account issues in gold standard judgements.

At its most base level, IR evaluation may be split into two schools of thought. Ranked retrieval evaluation and unranked retrieval evaluation. The most basic, yet most intuitive metrics for unranked retrieval evaluation are Precision, Recall and F-Measure. Precision is a measure of the number of relevant items returned against the total number of items returned, as defined in Equation 5.4.

\[
Precision = \frac{\text{RelRet}}{\text{Ret}}
\]

where \(\text{RelRet}\) is the number of relevant documents returned for a query, \(\text{Ret}\) is the total number of documents returned for the query.
Recall is defined as the ratio of relevant documents retrieved, as shown in Equation 5.5.

\[ \text{Recall} = \frac{\text{RelRet}}{\text{Rel}} \quad (5.5) \]

, where RelRet is the number of relevant documents returned for a query, Rel is the total number of documents in the collection that are relevant to a query.

Each of the metrics has their own advantages and applications. For example, in searching for patients eligible for clinical trials, users will require that all of the top ranking results are relevant to the information need, thus a high precision system is needed. An evidence based medical system, such as that which mines medical relations, would require high recall so that all information may be extracted. The F-measure, is the harmonic mean of the Precision and Recall metrics as in Equation 5.6.

\[ F_\phi = \frac{(\phi^2 + 1) \times \text{Precision} \times \text{Recall}}{(\phi^2)\text{Precision} + \text{Recall}} \quad (5.6) \]

The choice of \( \phi \) determines the weight given to Precision and Recall scores. For example, the \( F_1 \) score calculation which gives equal weighting to both metrics is shown in Equation 5.7.

\[ F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5.7) \]

A shortcoming of these metrics, is that they do not take into account the ranking of documents in the result lists. We will now discuss metrics used to evaluate ranked results.

**Mean Average Precision (MAP)**

*Mean Average Precision* (MAP) (Voorhees, 2000) is a highly stable metric that is often a standard measure for evaluations of ranked retrieval lists. The Average Precision for a single query is the average precision value of the top \( k \) documents returned. The MAP of a system is the average across all queries in the evaluation. More concretely, it is defined in Equation 5.8. The score falls between 1, best performance and 0, worst performance.

\[ \text{MAP}(Q) = \frac{1}{Q} \sum_{j=1}^{Q} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk}) \quad (5.8) \]
where \( Q \) is a set of queries, \( \{d_1, \ldots, d_m\} \) is the set of relevant documents, \( R_{jk} \) is the set of ranked retrieval results from the top result until you get to document \( d_k \).

**Precision At \( k \)**

In certain cases of IR evaluation, such as web retrieval, we may only be concerned with a small subset of documents that the user will view in the first page of results. MAP is insufficient in this case as it incorporates all retrieved documents. *Precision at \( k \)* allows us to focus on the precision of the top \( k \) documents. However, it is volatile as a measure and does not average well, as it is heavily influenced by the total number of relevant documents (Manning et al., 2008).

**R-Precision**

*R-Precision* is a measure that is highly correlated to Average Precision that aims to address the pitfalls of *Precision at \( k \)*. \( R \) in this case denotes the set of documents relevant to an information need. The metric is calculated through the precision of the top \( R \) documents.

**Binary Preference Score (BPref)**

The above metrics require a complete set of relevance judgements in order to give an accurate representation of performance. However, this is rarely the case. For example, in extremely large data collections such as web based systems, or even where relevance judgements require domain expertise leading to heightened expenses as in the medical domain. The *Binary Preference Score* is inversely related to the fraction of judged non-relevant documents that are retrieved before relevant documents (Buckley and Voorhees, 2004), in order to account for inadequate relevance judgements. On complete judgements, it has been shown to correlate highly with the MAP score.

### 5.3 Information Retrieval in the Biomedical Domain

Biomedical literature is an ever-growing dataset whose ambiguity poses problems for traditional IR approaches. Rather than being solely an issue of word sense disambiguation, effective biomedical search engines must be able to reason about the interactions and events e.g. mutations, translations, affecting entities in the texts. Secondly, if a search engine returns relevant documents a
biomedical researcher must still read the collection of documents retrieved and manually extract precise information. Though the retrieval of relevant records does speed up the process, manually mining the returned subset of literature is a time-consuming task. Therefore, methods have been developed to extract specific areas of text in order to shorten the time and effort required of end users. In this section, we provide an overview of systems and tasks developed to advance research in biomedical information retrieval.

**Query Expansion in the Biomedical Domain**

Early work in biomedical information retrieval focused on the use of query expansion in ad hoc retrieval. Initial experiments in Srinivasan (1995) highlight the importance of query expansion in alleviating issues of synonymy in biomedical literature. This was further expanded through the use of retrieval feedback, creating a more robust approach with improved results (Srinivasan, 1996). Aronson and Rindflesch (1997) explored expansion through the use of the UMLS Metathesaurus. While the authors found that their approach performed favourably, they concluded an optimal approach would combine UMLS and retrieval feedback based techniques. Further, given the rich synonymy and linguistic variation found in biomedical literature query expansion is an essential facet to effective biomedical IR (Stokes et al., 2009).

**Shared Tasks in Biomedical Information Retrieval**

Similar to the NLP domain, the growing interest in biomedical information retrieval led to the development of shared tasks which fostered much of the research in the burgeoning domain. The TREC Genomics track (2003-2007) provided the most substantial shared forum in advancing research in biomedical information retrieval. The corpora used for retrieval evaluation was composed of MEDLINE abstracts and full text HTML documents from the HighWire website. Though initial evaluations did focus on ad hoc retrieval evaluation, the track also featured domain-specific tasks such as text categorisation and question-answering. These tasks realise issues for biomedical researchers in searching across large amounts of text in identifying key information.

Text categorisation facilitates database curation whereby a text is deemed to fit into a certain category, allowing certain information to be extracted and placed into a structured database for analysis. Performing this task manually is time-consuming and costly. Through classification and further automation significant time is saved thus speeding up the analysis of genomic data.

\[^{2}http://ir.ohsu.edu/genomics/\]
Question-answering further reduces the need for researchers to read through large amounts of relevant data by extracting relevant answer passages from documents for an information need expressed in a natural language query, e.g. “What [GENES] make up the NFκB signalling pathway?”. Information extraction techniques are required to identify entities such as [GENES] in the text in order to correctly answer questions.

**Trends at TREC Genomics**

Throughout the years of the Genomics track, trends were noticeable. Initial experiments investigated the applicability of retrieval models. Results often showed Okapi BM25 (Billerbeck et al., 2004) to outperform language modelling (Ruiz et al., 2004) and tf-idf approaches. Given the high level of synonymy and variation in biomedical literature, research focused on concept based methods. For example, at TREC-13 Kraaij et al. (2004) investigated the usage of concept recognition and MeSH based expansions. This is continued throughout the track where template (Camous et al., 2005) and statistical methods (Aronson et al., 2005) were used in order to extract concepts. Concepts are then expanded using ontologies such as MeSH and UniProt as well as relevance-feedback methods.

**Post-TREC Methods to Biomedical Information Retrieval**

Following the discontinuation of the TREC Genomics track, research in the domain has continued, though at a slower pace. Work has continued in the extraction and expansion of concept terms in queries. For example, resources such as Wikipedia (Yin et al., 2010) have been implemented in expansions, seeing improvements in performance. As expansion methods may introduce further noise to retrieval results, concept re-ranking (Stokes et al., 2009) has seen the negative effects of expansion reduced.

In this section, we investigated the use of IR technologies in the biomedical domain. We have identified that query expansion plays a key role in alleviating issues related to the vast terminology difference in the domain. We have also seen the contributions of shared tasks in the area and the contributions of NLP systems in IR. In the next section, we will explore developments in IR in the clinical domain.
5.4 Information Retrieval in the Clinical Domain

In this section, we present an overview of existing methods and challenges for Information Retrieval in the clinical domain.

The storage, management and retrieval of clinical data using computerized methods is not a new research area. As early as the 60’s, there was research being conducted investigating the feasibility of using computers to manage data in a healthcare setting (Greenes et al., 1969), though much of this work focused on the use of structured methods (Mezzich et al., 1981; Sager et al., 1982). With a move towards unstructured methods, even in the early stages, concepts were a key priority given their high frequency as well as conveying much of the meaning of the text (Chute and Yang, 1992) as well as the relations between them (Hersh and Greenes, 1990).

Throughout the 90’s and early 2000’s, while there was interest in the development of Clinical Information Retrieval systems the research community suffered from the same data sharing issues that arose in the NLP community. As a result research focused on the needs of healthcare professionals (Hersh and Hickam, 1998) and discussions of in-house systems (Lee et al., 2006).

The cessation of the TREC Genomics track renewed interest in Clinical Information Retrieval, with several investigations into the PICO framework (Demner-Fushman and Lin, 2007; Boudin et al., 2010) for query formulation. However, these evaluations are performed on MEDLINE data, rather than clinical reports that are the focus of this thesis. The release of the BLULab dataset as described in Chapter 3 also opened avenues of research for the IR domain. Koopman et al. (2011b) explore the usage of ICD-9 codes, which identify concepts used for billing purposes, to create queries and relevance judgements. ICD-9 codes comprise of a code and textual component. The textual components are used to create queries, with the presence of an ICD-9 code in the documents meta-data used as a relevance judgement for the query. While novel, this approach produces few relevance judgements, and is not reliable as a standard.

5.4.1 TREC Medical Track

Since its inception in 1992, the Text REtrieval Conference (TREC) has provided an international forum for comparative evaluations of IR techno-
logies through annual challenges. Organisers supply participants with test collections, queries and relevance judgements, thereby providing participants the opportunity to evaluate and present their own approaches in the context of leading IR research. Perhaps uniquely to IR research, participants may submit manual or automatic systems. A manual system allows human interaction in the IR workflow, often at the query creation stage that may manipulate a query into a specific form for a system. Manual systems allow focus on retrieval performance without the need to focus on preprocessing errors. However, in a real-world application a user would not be expected to perform complex query manipulations. Therefore automatic systems feature no human interaction. As with other forms of text processing, IR systems must address a specific set of issues given a specific domain of text, be it medical, legal or web data. TREC facilitates domain-specific IR research through its numerous tracks that provide a dataset, queries and relevance judgements for a specific problem domain.

In 2011, TREC introduced the Medical track. The Medical track’s aim was to foster research on the retrieval of electronic health records based on the content of free-text fields. Techniques developed in this track identify cohorts suitable for clinical trials given a set of inclusion criteria specified through natural language queries. Electronic health records contain unstructured free-text describing a patient’s current medical status as well as tests, their history and their family’s history. As a result, the texts contain a considerable amount of negation, temporality and discussion of other individuals. Given that the criteria used to search are often highly specific, sophisticated methods are required to correctly mine and reason about the individuals in the report, before returning a relevant document.

An important issue that has arisen for NLP and IR to overcome in the clinical domain, is the sensitive nature of the data. It is key that a patient’s identity cannot be discovered, therefore several security and ethical measures must be enforced to safeguard the patient’s privacy. In the past, methods of de-identification either were not available, or not good enough for healthcare institutions to willingly distribute data, leaving researchers to focus on other sources of non-clinical, medical data such as MedLine articles. More recently de-identification systems as well as security and legal measures have been developed to ensure privacy of individuals. As a result, data can now be distributed subject to license agreements and ethics approval.
Overview of the document collection

The dataset used in the TREC Medical Track (both 2011 and 2012) was one of the largest distributions of de-identified medical data. The dataset, comprising of 101,711 clinical reports was obtained from the BLULab group’s NLP Repository at the University of Pittsburgh. The unit of retrieval for the task is a visit. A visit contains one or more clinical reports that refer to a specific admission in a patient’s history. TREC provided a mapping file outlining the mapping between the 17,266 visits and reports they contain.

Queries (or topics) provided for this task specified a set of inclusion criteria in order to identify cohorts. Inclusion criteria may specify demographic information, such as age, gender etc. as well as details required in a patient’s medical history, for example a treatment or specific condition. In 2011, 35 queries were provided with 50 new queries provided for the 2012 task.

In the following subsections we will present an overview of the TREC Medical track in 2011 and 2012, highlighting the performance and variability of the submitted systems.

5.4.1.1 Medical Track 2011

In 2011, TREC included a medical track for the first time. Given that it was a domain containing relatively little research, TREC medical track 2011 represents an epoch in Clinical IR. As a result methods proposed at the 2011 Medical Track were exploratory with a lot of variance and differences between the systems implemented. In this section, we present an overview of the systems submitted to TREC Medical Track 2011 and assess the impact of different strategies. The approaches used in systems are summarised below:

- **QE**: Query Expansion is a commonly used approach in IR to address issues caused by synonymy, or near synonymy through adding more terms to an initial query. For example, if a user uses a highly specific terminology, query expansion may be employed to improve recall by using more general terms.

- **Sem**: Semantic resources may be used to perform concept normalisation. With concept normalisation, concepts in the text are mapped to a unique identifier. Leveraging this information aims to minimise ambiguity in the user’s information need.
• **lVar**: Similar to query expansion, lexical variants of a token are generated and added to a query, rather than tokens that are related through co-occurrence statistics or semantic relations.

• **Gen**: As clinical queries specify demographic information, such as gender. Systems may filter documents so that gender in the document will match the gender in the query.

• **Age**: In this case filtering is based on the patient’s age in the visit and that found in the query.

• **CI**: Concept Identification identifies ontology concepts e.g. UMLS, in the report text and in the query.

• **Neg**: Negation detection indicates if a query term occurs within a span in the text that is negated.

• **Assert**: This extends negation detection to further Assertion Status. A medical condition may have occurred in the past, or affect a family member.

• **Rel**: Relation detection identifies relations between concepts, such as the relationship between a treatment and condition.

• **Prox**: Proximity matching ensures that certain lexical items co-occur within a certain span. For example, a treatment and condition or a condition and its specifier such as *Acute myocardial infarction*.

• **Weight**: Query weighting strategies add additional weight to certain, important query terms. Thus biasing the ranking to ensure that documents that contain the weighted term appear toward the top of the ranked list.

• **Sect**: Clinical reports are divided into logical sections. Systems may aim to manipulate this information to only check certain sections for query terms.

Systems submitted to TREC Medical Track 2011 who published system reports are presented in Table 5.3. Systems in Table 5.3 are ordered from top to bottom in order of performance on the Binary Preference score metric at the TREC 2011 Medical track.

**The role of Query Expansion at TREC Medical 2011**

*QE is a common trend among the submitted systems with all but six groups utilising some form of expansion, highlighting the need to address the rich level of synonymy found in clinical texts.* Among those groups that used QE, there was variation in the resources used to perform expansion. Medical specific ontologies such as the UMLS or MeSH (Daoud et al., 2011) were commonly used.
along with web based resources like MedLine (Cordoba et al., 2011). Further approaches investigated the usage of non-domain specific ontologies such as DBpedia (Karami et al., 2011) or resources such as Wikipedia. One system (Ozturkmenoglu and Alpkocak, 2011) investigated the generation of lexical variants (lVar), showing an improvement in performance when combined with other semantic techniques. Meta-data in the text e.g. ICD-9 codes were used by a number of systems, however its contribution to the task was largely non-uniform, often specific to particular queries, rather than to the problem as a whole, with a high level of human interaction (Bedrick et al., 2011) in the manual runs with automated approaches noting negligible impact in the usage of ICD-9 information (He et al., 2011).

The use of \textit{Sem} was the most common approach used across all systems, again recognising the need for normalising medical textual data given its rich synonymy and lexical variations. With concept normalisation, concepts in the text are mapped to a unique identifier known as a \textit{CUI}. Leveraging this information aims to minimise ambiguity in the user’s information need. It has yielded some positive results in concept-based retrieval methods (Demner-Fushman et al., 2011) that required manual generation of queries due to their high complexity. Shortcomings of this approach include the requirement for human intervention along with ambiguity that may be introduced through using the ontology. Though it is a promising approach it requires further work to remove complexities for the user base.

Daoud et al. (2011) note very few queries requiring information on age. Despite this, filtering was found in the three top performing automatic runs (King et al., 2011; Gurulingappa et al., 2011; Goodwin et al., 2011).

**Concept Identification at TREC Medical 2011**

Methods employing conceptual methods required a system to identify concepts (CI) in the text. These methods ranged from simplistic greedy algorithms matching against dictionaries (Eichmann, 2011) to purpose built NER systems (Cordoba et al., 2011). \textit{Standalone performance evaluations of many of these systems are not provided in the TREC Medical Track 2011 literature.}

**Assertion Classification**

Negation detection (\textit{Neg}) and Assertion Classification (\textit{Assert}) were a common trait among top performing systems where 6 out of 7 approaches implemented \textit{(Neg)} and/or \textit{(Assert)}. One system (Gurulingappa et al., 2011) employed relation detection (\textit{Rel}) with little impact on retrieval performance.
<table>
<thead>
<tr>
<th>System By</th>
<th>QE Sem Gen Age Prox CI Neg lVar Assert Rel Weight Sect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Automatic</strong></td>
<td></td>
</tr>
<tr>
<td>King et al. (2011)</td>
<td>x x x x x x x x</td>
</tr>
<tr>
<td>Gurulingappa et al. (2011)</td>
<td>x x x x x x x</td>
</tr>
<tr>
<td>Goodwin et al. (2011)</td>
<td>x x x x x</td>
</tr>
<tr>
<td>Zhu and Carterette (2011)</td>
<td>x</td>
</tr>
<tr>
<td>Schuemie et al. (2011)</td>
<td>x x x x x</td>
</tr>
<tr>
<td>Limspatham et al. (2011)</td>
<td>x x x x x</td>
</tr>
<tr>
<td>Demner-Fushman et al. (2011)</td>
<td>x x x x x</td>
</tr>
<tr>
<td>Ozturkmenoglu and Alpkocak (2011)</td>
<td>x x</td>
</tr>
<tr>
<td>Koopman et al. (2011a)</td>
<td>x</td>
</tr>
<tr>
<td>Bhattacharya et al. (2011)</td>
<td>x</td>
</tr>
<tr>
<td>Dinh and Tamine (2011)</td>
<td>x</td>
</tr>
<tr>
<td>Cordoba et al. (2011)</td>
<td>x x x x x</td>
</tr>
<tr>
<td>Wu et al. (2011)</td>
<td>x x x x x</td>
</tr>
<tr>
<td>Wu and Fang (2011)</td>
<td>x x</td>
</tr>
<tr>
<td>Eichmann (2011)</td>
<td>x x x x x</td>
</tr>
<tr>
<td>Karami et al. (2011)</td>
<td>x x x</td>
</tr>
<tr>
<td>Daoud et al. (2011)</td>
<td>x x x x x</td>
</tr>
<tr>
<td>He et al. (2011)</td>
<td>x x x</td>
</tr>
<tr>
<td>Gobeill et al. (2011)</td>
<td>x x x</td>
</tr>
<tr>
<td><strong>Manual</strong></td>
<td></td>
</tr>
<tr>
<td>Demner-Fushman et al. (2011)</td>
<td>x x x x x</td>
</tr>
<tr>
<td>Zhang et al. (2011)</td>
<td>x x</td>
</tr>
<tr>
<td>Dinh and Tamine (2011)</td>
<td>x</td>
</tr>
<tr>
<td>Gurulingappa et al. (2011)</td>
<td>x x x x x x</td>
</tr>
<tr>
<td>Cogley et al. (2011)</td>
<td>x x</td>
</tr>
<tr>
<td>Wu et al. (2011)</td>
<td>x x x</td>
</tr>
<tr>
<td>Bedrick et al. (2011)</td>
<td>x x x</td>
</tr>
<tr>
<td>Toldo and Scheer (2011)</td>
<td>x</td>
</tr>
</tbody>
</table>

Table 5.3: Approaches used at TREC Medical Track 2011
The impact of Structural Information

Two systems used proximity information (Daoud et al., 2011; Wu and Fang, 2011) in their systems. Proximity information ensures that concepts identified in the text are used to retrieve documents where the concepts occur within a certain span of one another, for example, a medical condition and a treatment co-occurring. These approaches were shown to improve retrieval performance in both systems, however these approaches were not comparable to top-performing systems.

*The use of document structure (Sect) had a non-uniform impact on results* as it is leveraged in relatively high scoring (Schuemie et al., 2011) and low-scoring systems (Karami et al., 2011).

The choice of retrieval model

Varying models of retrieval were used in the task ranging form standard tf-idf and Okapi methods to language modelling techniques, LSI (Latent Semantic Indexing) and Concept-based methods. While each approach had its merits, there was no approach that stood out as an obvious state-of-the-art. For example, though King et al. (2011) had the top performing automatic run through filtering based on demographic information and negation detection, Demner-Fushman et al. (2011) attained the top performing manual run using a conceptual approach that required manual construction of queries. One could posit that automating a conceptual approach may outperform current systems.

To conclude, the problems of synonymy and variation that are commonplace in medical records must be addressed by the system, either through query expansion or normalisation techniques. Top performing systems filtered results based on demographic information such as age and gender. These systems also accounted for negation and assertion statuses of medical conditions found in the text. These techniques highlight the domain-specific difficulties in Clinical IR that cannot be addressed by simpler keyword approaches.

5.4.1.2 Medical Track 2012

The TREC Medical Track 2012 followed the same format of the previous year’s track. The test collection remained the same, with data obtained from the University of Pittsburgh. The task aim remained to identify cohorts who fit a set of inclusion criteria specified in natural language queries. 50 new queries were created in order to evaluate systems. Given the high level of similarity between
the two years, systems submitted in 2012 aimed at honing, fine-tuning and extending previous systems while performing much more investigative analysis in order to determine methods that aid retrieval performance. A breakdown of the features used in the retrieval systems is presented in Table 5.4. All systems in Table 5.4 are ranked according to official results at TREC 2012 using the Binary Preference score metric.

**Query Expansion at TREC Medical 2012**

Similar trends continued in the 2012 track as Query Expansion (QE) and extensive use of semantic resources (Sem) addressed the usage of non-standard terminology in clinical texts. However, the methods to perform query expansion varied throughout the submitted systems. (Koopman et al., 2012b) extended their bag-of-concepts approach through including more relationships, rather than simply synonymy. The nature of relationships between concepts e.g. *is-a, part-of* determined the weight a concept was given, an approach that improved performance. Martinez et al. (2012) also saw an improvement in performance using the Personalized PageRank algorithm. The attitudes to refining query expansion for the domain are best summed up in (Amini et al., 2012) and (Bedrick et al., 2012) where it was shown that haphazard query expansion can introduce a significant amount of noise that can be reduced through more novel expansion methods and the use of stop-lists for concepts.

**The impact of Assertion Classification**

After the prominence of negation detection and other forms of information extraction at the 2011 Medical Track, researchers investigated whether these techniques really contribute to the task. (Diaz et al., 2012) found that negation didn’t necessarily improve systems due to the low number of terms affected by negation in the text as well as very few negated terms appear in the test queries. Despite this, a number of systems (Limsopatham et al., 2012; Amini et al., 2012) found assertion classification, primarily negation detection to be beneficial to the retrieval task. This is in keeping with findings that lack of assertion classification can lead to a number of false negatives (Wu et al., 2012; Edinger et al., 2011). Although general assertion classification did not greatly improve performance at the 2012 Track, 5 of the top 6 systems all implemented negation detection.

Zhu and Carterette (2012) achieved top performance in the automatic systems through the combination of multi-level evidence. The system queries both report-based indices and visit-level indices. Relevancy judgements are then
combined using the ComboSUM method, i.e. the summation of the set of similarity values for the report and visit indices. It also highlights the use of diverse resources such as Wikipedia for query expansion along with proximity matching. The second top-ranking system, Qi and Laquerre (2012) highlighted that negation handling may not be enough as further assertion status such as Possible or Conditional as described in Chapter 4. The top ranking manual system submitted by Demner-Fushman et al. (2012) also outperformed automatic approaches, incorporating manual query expansion with the UMLS as well as assertion classification and demographic filtering.

<table>
<thead>
<tr>
<th>System By</th>
<th>Sem</th>
<th>Neg</th>
<th>CI</th>
<th>QE</th>
<th>Assert</th>
<th>Age</th>
<th>Gen</th>
<th>Prox</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Automatic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhu and Carterette (2012)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Qi and Laquerre (2012)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tinsley et al. (2012)</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Callejas et al. (2012)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limspopathom et al. (2012)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Martinez et al. (2012)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Miao et al. (2012)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Koopman et al. (2012b)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamdan et al. (2012)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demner-Fushman et al. (2012)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Leveling et al. (2012)</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bui et al. (2012)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goodwin et al. (2012)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amini et al. (2012)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wu et al. (2012)</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diaz et al. (2012)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bedrick et al. (2012)</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Manual</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demner-Fushman et al. (2012)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Bedrick et al. (2012)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.4: Approaches used at TREC Medical Track 2012

In conclusion, TREC 2012 Medical Track facilitated further research into domain-specific methods that aim to improve the performance of medical information retrieval platforms. These methods include assertion classification, novel query expansion using diverse resources and concept-centric retrieval methods, however none have evaluated the performance application of a machine learning based MER system.

Post-TREC Medical Track

Though the Medical track was discontinued in 2012, the initial release of the dataset and relevance judgements as part of the track has facilitated post-
submission experiments by participants. Much of this work has investigated exploiting semantic information to determine a relevant document (Koopman et al., 2012a). Babashzadeh et al. (2013) investigate the usage of the UMLS Metathesaurus to determine the relatedness of concepts in the query and in documents. The combination of this system with relevance feedback methods saw minor improvements on the state of the art. The combination of textual and conceptual retrieval has also seen gains in performance (Limsopatham et al., 2013).

5.5 Summary

In this chapter we explored the theoretical background of Information Retrieval in the clinical domain as well as a discussion of research in the area. We will now summarise our discoveries with respect to our research aims as outlined in the introduction to this chapter.

What model is best suited to the retrieval of documents in the clinical and biomedical domain?

In Section 5.2.1 we presented an overview of the information retrieval models available in retrieving a ranked list of documents, given an information need represented as a textual query. We highlighted that Vector Space models are inadequate as they take no account of document length and term dependencies. Probabilistic methods such as the Okapi BM25 makes a better case for inclusion as it accounts well for document length, however it too does not account for term dependencies. Language modelling gives us the capacity to include term dependency information, but falls short in that it cannot leverage structural information such as that provided by the NLP techniques described in Chapters 3 and 4. The most suitable model for the approach undertaken in this thesis combines Language Modelling and Inference Network methodologies. Language modelling allows for term dependencies while Inference Networks provided a structured framework and query language to utilise information extraction techniques.

How are Information Retrieval systems evaluated?

Section 5.2.2 provided descriptions of the most commonly used evaluation metrics in Information Retrieval research. For the evaluation of unranked retrieval, we see the basic metrics of Precision, Recall and F-measure may be used. For the evaluation of ranked retrieval methods, MAP, RPrec and Pre-
cision at k are all commonly used. However, these three metrics require relevance judgements that are comprehensive and are missing no documents. Given the high level of expertise required in annotating clinical documents, there is a high possibility that some documents are not annotated. Therefore, the key metric investigated in this thesis will be the BPref metric, as it is robust to such issues while still correlating well with other metrics.

What problems arise in clinical and biomedical information retrieval systems?

From the literature we can see that synonymy, linguistic variation and ambiguity are key issues throughout information retrieval in both the clinical and biomedical domains. Non-standardised use of hyphenation in gene names has been a consistent issue in the biomedical domain in both NLP and IR. In clinical reports, entities are further complicated as they may be modified in a number of ways e.g. Query term CAD may occur in a document, but it may be affecting another individual deeming it irrelevant, e.g. Father positive for CAD.

What strategies have been used in the literature to address domain specific issues?

Throughout shared tasks a multitude of methods have been suggested in overcoming barriers in biomedical and clinical information retrieval. Query expansion and lexical variant generation using both ontology and relevance feedback methods are commonly implemented. The extraction of concepts and other semantic information, rather than keywords, from texts is an area attracting huge interest. If we retrieve concepts rather than terms, this may eliminate problems of mismatching. However, if concept extraction is not accurate it may do considerably more harm to Information Retrieval performance than keyword retrieval.

To what end have information extraction techniques been applied to information retrieval in the biomedical and clinical domains?

Several approaches in the literature have highlighted the use of techniques such as Medical Entity Recognition and Assertion Classification using both rule-based and machine-learning methods. Very few system descriptions feature standalone evaluations of these information extraction techniques, or how they are implemented in a retrieval setting. As a result, it has previously been difficult to determine the impact of natural language processing techniques in Information Retrieval applications and the required performance of NLP techniques in order to have a positive impact on IR effectiveness.
In the next chapter, we build on our findings in this chapter to develop a clinical information retrieval system that leverages NLP techniques in automatic query formulation and retrieval.
CHAPTER
SIX

IMPLEMENTING A CLINICAL IR APPROACH

6.1 Introduction

In this chapter, we describe the challenges of designing a Clinical Information Retrieval approach that leverages information extracted through Medical Entity Recognition, Assertion Classification and Concept Normalisation systems. The previous chapter provided a discussion of retrieval models available to the community, with a language modelling and inference network approach deemed the most suitable due to its extensible query language that can leverage complex annotations as well as its ability to handle term dependencies. The implementation we shall use is Indri (Strohman et al., 2005), first discussed in Chapter 5 and will be described in further detail in Section 6.3. In this chapter we investigate the following:

- The functionality and features of the Indri Retrieval system.
- The integration of NLP techniques with the Indri Retrieval system.
- The development of tools used in automatically translating natural language queries into Indri’s structured query language.
- The data and query sets used in evaluating the Clinical Information Retrieval System

In Section 6.2, we present an overview of the datasets used to evaluate the clinical Information Retrieval system. Section 6.3 discusses the Indri Retrieval system. In particular, we provide an overview of aspects of Indri’s query language
that facilitate the usage of Natural Language Processing approaches. Section 6.4 describes tools developed in order to automate the process of querying such that users need only present natural language queries to the system, which are translated into a structured query language as well as performing query expansion. In Section 6.5, we provide an overview of the clinical Information Retrieval system’s workflow.

Figure 6.1: Diagram of Information Retrieval System and Datasets

6.2 Experiment Dataset

To evaluate the performance of the clinical IR system, we use the query sets, relevance judgements and document collection from the TREC 2011 and 2012 Medical Tracks. Utilising this dataset allows a fair comparison of the work performed in this thesis to past and future work in the community. The TREC Medical track task requires systems to identify cohorts eligible for clinical trials given a set of inclusion criteria specified through a natural language query. First, we will provide an overview of the document collection, followed by a discussion of the query set.

6.2.1 TREC Medical Track Document Collection

The same dataset was used in both the 2011 and 2012 TREC Medical Tracks. It was obtained through the University of Pittsburgh’s BLULab NLP Repository.
This is the same dataset as discussed in Chapter 4. It is comprised of 101,711 de-identified clinical reports. There are eight report types in the dataset, as shown in Figure 6.2.

The most numerous are Radiology Reports (RAD), followed by History and Physical Examinations (HP), Emergency Room reports (ER), Discharge Summaries (DS) and Progress notes (PGN). Surgical Pathology (SP) and Echocardiogram (ECHO) reports make up just under four percent of the distribution. The report types vary in content, structure and length. RAD and ECHO reports are typically short reports, often containing numerical content or statement of procedures. SP reports are often information dense, though longer than RAD and ECHO reports, providing a discussion of surgeries and providing references for the analysis of procedures. HP, ER, DS and PGN reports are discussion based. They represent a monologue of the physician in analysing a patient’s state, previous admissions and history, as well as diagnosis or prescription. It is these types of report that most require some form of NLP for information extraction as a lot of pertinent information is discussed in verbose natural language.

As these reports often describe the same patient, just at different stages of their hospital visit, the unit of retrieval for the TREC Medical track is a visit. A visit combines all reports relating to a patient’s particular visit into one document. Therefore, a given visit document may contain several documents of any report types, e.g. an ER report details the admission, followed by 3 RAD reports specifying X-ray tests, followed by DS report giving diagnosis and prescribed treatment. TREC provided a mapping file outlining the mapping between the
17,266 visits and reports they contain. All runs described in this thesis use a pre-processing concatenation approach. First, visit documents are created through the concatenation of reports according to the mapping file. Indices are then built using this new visit collection.

### 6.2.2 TREC Medical Track Query Sets

The queries used to evaluate the clinical IR system come from the TREC 2011 and 2012 Medical tracks. The 2011 set contains 30 queries, 29 of which are used in evaluation. The 2012 set contains 50 queries, 49 were used for evaluation purposes. The queries left out, were also left out at the respective tracks as the relevance judgements were not adequate for these queries. The queries from the task specify a set of inclusion criteria to specify patients eligible for a clinical trial. Inclusion criteria may describe a patient’s current medical condition, treatments undergone or tests prescribed for them. For example, in Query 129 below, we require patients currently experiencing chest pain and have undergone the test CT angiography.

- **Query 129**: Patients admitted with *chest pain* and assessed with *CT angiography*.

Aside from medical information, the inclusion criteria may also discuss a patient’s demographic information, such as age, gender etc. For example, Query 109 requires female patients and Query 136 specifies children.

- **Query 109**: *Women with osteopenia*.
- **Query 136**: *Children with dental caries*.

In the next section, we provide an overview of the Indri retrieval system, in particular its query language that provides the basis for the clinical IR system in this thesis.

### 6.3 Indri Retrieval System

The treatment of documents as unordered sets of term frequencies, known as a *bag-of-words*, is a common pillar of several IR systems. Early IR implementations such as the Vector-Space Model (Salton et al., 1975) and the probabil-
istic Okapi (Robertson and Sparck-Jones, 1976) that use a bag-of-words approach provide the basis for most modern retrieval approaches such as the BM25 model (Robertson et al., 1994). Though these approaches post impressive results, a wealth of information, including word-order, document structure, deeper semantics and term-dependencies are discarded. Clinical texts contain a rich sub-language (Sager, 1981) that may impact the performance of standard bag-of-words IR approaches.

Language models (Hiemstra, 2000), or n-gram language modelling, provide a technique to account for term-dependencies. These models rank documents according to the likelihood that a document could generate the given query. For \( n \geq 1 \), the sequential nature of the text may be incorporated. However, language modelling based methods only use text-based information, i.e. the terms, where we require the ability to manipulate named entities, assertions and demographic information extracted from the text.

Indri’s (Strohman et al., 2005) retrieval model combines a language modelling approach (Song and Croft, 1999) with that of an inference network model (Turtle and Croft, 1991). The joining of both approaches accounts for the sequential nature of text, term-dependencies as well as a model that can generate an “understanding” of documents and queries (Turtle, 1991; Metzler, 2007).

The Lucene project\(^1\) is perhaps Indri’s closest rival in terms of functionality, performance and the fact they are both open-source and widely used in the research community. Evaluation experiments (Turtle et al., 2012) have shown Indri outperform in terms of throughput and ranking results in shared task data-sets such as TREC. Furthermore, though both systems implement field-based retrieval, Indri’s has been shown to be more robust and feature rich. These gains provide us the ability to seamlessly integrate NLP and IR approaches.

### 6.3.1 Indri Query Language

A key aspect of the Indri Retrieval system is its robust and comprehensive query language. The Indri Query Language, based on INQUERY syntax (Callan et al., 1992), provides the ability to perform phrasal matching, weight terms, annotate synonyms and field-based retrieval. In this section, we will discuss the term operators that will facilitate its use in conjunction with NLP

\(^1\)http://lucene.apache.org/solr/
systems.

6.3.1.1 Automatically Translating Queries

In this section, we describe the aspects of the Indri Query language that are used to leverage information generated through NLP techniques. All of the following features of the queries are automatically generated through the Clinical Information Retrieval system, thus removing requirements of training for healthcare professionals.

Stopping and Stemming

The first task to perform is the removal of stopwords and the reduction of terms to stems, for both queries and the document collection. A list of stopwords was compiled from generic English stopwords as well as high frequency terms in the corpus (Bui et al., 2012), e.g. *presented*, . Stemming was performed using the Porter stemming algorithm (Porter, 1980). While there are considerations that stemming may hamper performance through information loss, e.g. *AIDS* may be stemmed to *AID*, experimentation in the area has shown that this possible performance issue is outweighed by its generalisation ability (Martinez et al., 2012).

Facilitating Keyword Searching

The most basic form of search in the Indri Query Language is the #combine term operator. Term operators correspond to belief nodes described in Section 5.2.1.4. In using this operator, all terms it encloses are given an equal weighting when scoring a document. Its usage is shown in Query 1. below.

1. #combine( Patients with acute vision loss secondary to glaucoma )

With the application of stopword removal, the query becomes Query 2 below:

2. #combine( acute vision loss secondary glaucoma )

However, as vision loss is a secondary condition to glaucoma, we may want to assign more importance to glaucoma. To do this we may use Indri’s #weight operator, rather than #combine, as in Query 3.
3. \#weight( 5.0 acute 5.0 vision 5.0 loss 3.0 secondary 10.0 glaucoma )

Phrasal Search

As previously discussed in Chapter 5, a shortcoming of keyword search is it takes no account for term dependencies. To account for term dependencies, we may use the \#N operator where \(N\) is an integer representing the allowable number of tokens between terms enclosed within it. For example, in Query 4 the \#1 operator is used, indicating that vision and loss must occur within 1 token of each other, in that exact order.

4. \#combine( acute \#1( vision loss ) secondary glaucoma )

The \#uwN operator allows the ordering between terms to be variable within a span of \(N\). The use of this operator rather than \#N, allows for more variation in phrases i.e. loss of vision, vision loss etc. Its usage is shown in Query 5 below:

5. \#combine( acute \#uw2( vision loss ) secondary glaucoma )

Allowing for Query Expansions

To further account for linguistic variation in the document collection we may employ the usage of query expansion. Query expansion may be implemented in Indri using the \#syn operator. Its usage is demonstrated in Query 6. The \#syn operator gives equal weighting to all phrases or terms it encapsulates. If the user wishes to assign different weights to expansions the \#wsyn operator may be used.

6. \#combine( acute \#uw2( vision loss ) secondary \#syn( glaucoma \#1( ocular disorder ) ) )

Field-Matching

Field-matching requires the system to look in specific spans of a document for term matches. Spans, or fields in a document are marked using an XML-like
mark-up e.g. `<AGEGROUP>child</AGEGROUP>`. These fields can then be queried for specifically by using the format `term.field` as shown in Query 7.

```
7. #combine( child.agegroup #1(dental caries) )
```

**Filtering search**

The final type of operator we will discuss are filtering operators. These operators, `#scoreif` and `#scoreifnot` ensure that only documents that contain an exact term, phrase or field are returned. The area enclosed by the filtering operators are divided into two components. The first specifies the criteria which documents must match in order to be relevant to the query. The second component contains a typical query used to score and rank relevant documents. An example is shown in Query 8.

```
8. #scoreif( child.agegroup #combine( #1(dental caries) diabetes ) )
```

For a document to be relevant to Query 8 it must first contain the field `agegroup`, which must have the value `child`. If this criteria is met, the document is then scored according to the combination of the phrase `dental caries` and the term `diabetes`.

### 6.3.1.2 Calculating a Ranking Score for a Document

The Indri Retrieval model combines language modelling and inference network approaches. For each query, an inference network is created. Using this inference network we may compute $P(I = 1 \mid D)$, the probability that $D$ satisfies the information need $I$. To perform this, we first compute $P( qr \mid D )$, where $qr$ is the representation node e.g. the phrase “liver failure” occurred, for each of the representation nodes in the query. Its calculation is shown in Equation 6.1.

$$
P(qr|D) = \frac{tf_{qr,D} + \mu P(qr|M)}{|D| \mu}
$$  \hspace{1cm} (6.1)

where $tf_{qr,D}$ is the term frequency of $qr$ in $D$, $P( qr \mid M )$ is the expected value of the representation node $qr$ in the model of the entire collection $M$. $\mu$ is a tuning parameter for smoothing.

Once $P( qr \mid D )$ is calculated for all representation nodes in the query, they may
be combined using the belief operators such as \#combine and \#weight. A full
description of their calculation is described in Metzler (2007). Following the
combination of beliefs, we will have calculated $P(I = 1 \mid D)$, thereby allowing
the ranking of documents relevant to a query.

In the next section, we will describe the tools and methods used in conjunction
with the query language to create the Clinical IR system.

6.4 Tools for retrieval

In this section we describe the tools implemented to build on Indri’s core func-
tionalities to facilitate Clinical IR. The use of these tools allow the user to enter
a natural language query, while still having the ability to leverage Indri’s struc-
tured query language.

6.4.1 Query Expansion Methods

Query expansion is a vital aspect of Clinical Information Retrieval. However,
there are several issues that one must be aware of in expanding query terms.
For example, if an expansion is too broad it may introduce noisy results. If we
introduce too many expansions, the original focus of the queries may be also
lost. In this section, we present two tools developed as part of this thesis to
investigate the effects of query expansion. We leverage ontology based expan-
sion in this work rather than feedback based methods in order to capture the
high level of language variation found in clinical documents.

6.4.1.1 MeSH Expander

The first tool we will discuss is the MeSH expander. This tool is based on the
PubMed search facility’s ability to map keywords to MeSH terms, then to re-
lated keywords. First, concept spans are identified using the system discussed
in Chapter 3. The extracted concept spans are then passed to PubMed for ex-
pansion using the Entrez eutils service (Sayers, 2013). The returned translated
query is then parsed, and expanded terms are extracted. For example, if we
pass the span colon cancer to PubMed, we retrieve the translated query shown
below:

- "colonic neoplasms"[MeSH Terms] OR ("colonic"[All Fields] AND "neo-
plasms"[All Fields]) OR "colonic neoplasms"[All Fields] OR ("colon"[All Fields]
The spans enclosed in double quotes are used as expansions. Items in squared brackets denote the field in which the term is searched, and are disregarded for the purposes of expansion.

### 6.4.1.2 UTS Expander

The second approach to expanding queries uses the UMLS Terminology Services (UTS). With the UTS, a concept span may be mapped to a CUI in the UMLS. After assigning a CUI to the concept, we may exploit this information to mine relationships between CUI’s in order to create expansion terms. The algorithm for generating UTS expansions for a given concept span is outlined in Algorithm 3 and is free to download \(^2\).

#### Algorithm 3 Generating Expansions through the UTS for a given concept

```plaintext
CONCEPT ← INPUT
EXP ← {}
matchingCUIS ← (CONCEPT→GetCUIS())
for all cui in matchingCUIS do
    EXP ← push(cui→text)
    relations ← (cui→GetRelations())
    for all rel in relations do
        if (isAllowedRel(rel) == true) then
            relCUI ← (rel→getCUI())
            EXP ← push(relCUI→getText())
        end if
    end for
end for
return EXP
```

After first identifying a matching CUI for a given concept, we add the textual component of the CUI to the list of expansions for the term. Following this, we gather all the possible relations between the given concept and other CUIS in the network. For each relation we then determine if it is an allowable relation. If it is an allowable relation, we extract the related concept’s text and add it to the list of expansions.

The check in place for allowable relations give us granularity in our expansion process. The possible relations between concepts in the UMLS Metathesaurus are given in Table 6.1.

\(^2\)https://www.github.com/mccogley/UTS_Expander
<table>
<thead>
<tr>
<th>Relation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RB</td>
<td>Broader Relationship</td>
</tr>
<tr>
<td>RL</td>
<td>Similar or Alike concepts i.e. synonymous</td>
</tr>
<tr>
<td>RN</td>
<td>Narrower Relationship</td>
</tr>
<tr>
<td>RO</td>
<td>Relationship other than broader, narrower and synonymy</td>
</tr>
</tbody>
</table>

Table 6.1: Initial Allowed Relations between Concepts in the UMLS

In its unrestricted form, the UTS expander allows relations of all types presented in Table 6.1. However, the user may tune the granularity of expansions through the exclusion of certain relationship types in the allowable relations rule. For example, introducing RB would aim to improve recall, though at the greatest likelihood of introducing noise. In Chapter 7, we will investigate the usage of unrestricted expansions and the impact of refining the relation type.

### 6.4.2 Gender and Age Detection

As discussed in the previous section, a query may specify an inclusion criteria that includes the required patient’s medical history, as well as demographic information such as age or gender. In this section we describe methods in the application and use of age and gender information for querying.

#### Age Detection

The first stage in allowing the leveraging of age information in querying is to annotate the documents with the required fields. As reports are de-identified, age information is generalised. For example, the sentence “The patient is a 24 year old male” is de-identified to “The patient is a **AGE[20s] male”. This generalised form provides a simpler solution to extracting a patient’s age, as it can be extracted with a simple regular expression. For example the Perl regular expression, `/\*\*AGE\[(^\]+)\]/`, is used to extract age information in the text. Once we have extracted the patient’s age, we create two new fields in the document, AGE and AGEGROUP.

AGE is a numeric field which specifies the age of a patient in a document. So for the above example the AGE field becomes, `<AGE >20</AGE>`. The AGEGROUP field classifies patients into three categories, child which is any individual under 18, adult within the age range of 18 to 65 and elderly for anyone aged over 65. Given the example above, the AGEGROUP field becomes `<AGEGROUP>adult </AGEGROUP>`.

Figure 6.3 shows the distribution of age groups among the visit documents.
The child group is by far the minority, as the adult population takes nearly 60\% of the distribution. Elderly patients make up just over 20\% of the distribution with 18\% containing no age information in the text (unknown).

With fields in the document collection now annotated, we may perform field matching in the queries. A set of rules and regular expressions were developed to extract age information in queries and create field-matching terms. For example, adults under 60 becomes \#combine( adult.agegroup \#less( age 60 ) ) where the \#less operator indicates to retrieve documents where the field age has a value less than 60.

**Gender Detection**

Following age detection, we integrate gender detection. As there are no fields available in the text to extract gender information as for the age of the patient, a search over the document is required to determine the gender of the patient. The search performs rule-based and regular expression based checks to determine gender. For example, documents often contain the pattern Patient is a [ AGE ] [ GENDER ] with [ PROBLEM ]. Checks for these patterns, as well as keyword searches, such as for gender pronouns are performed. Once the search has completed, the system aggregates the information and adds a gender field to the text, whose value is the gender of the patient in the document, e.g. <GENDER>female</GENDER>.

In performing the annotation, we arrive at a gender distribution as shown in Figure 6.4. The distribution is slightly skewed in favour of female patients and 14\% of documents contain no explicit gender information in their text (unknown).
A series of rules were created to map queries to the field-based matching method such that terms such as *women* become female.gender.

### 6.4.3 Medical Entity Recognition and Concept Normalisation

A key aspect of this thesis is an investigation of applying NLP technologies from shared tasks in the domain to a clinical Information Retrieval workflow. In this section, we focus on Medical Entity Recognition and Concept Normalisation.

**Medical Entity Recognition**

Medical Entity Recognition requires the identification of a span of text that contains a medical concept, and to classify the type of concept of the span. For our Clinical IR system, we explore two methods of utilising this information.

The first method uses Indri’s field matching technique. The i2b2/VA MER system as discussed in Section 3.3.2 is used on both the document collection and query set. In the document collection the system is used to annotate XML like spans in the text, e.g. `<PROBLEM>diabetes</PROBLEM>`. For queries, it adds the requirement for field matching, e.g. diabetes.problem. Three concept types, from the i2b2/VA challenge, are annotated: *Problem, Treatment* and *Test*.

The second approach simply uses the i2b2/VA MER system to annotate phrasal spans in the query text. This removes the need for the user to perform keyword search, as well as the need to process the document collection. For example, given the keyword query `#combine( colon cancer )` the system annotates the phrasal boundary `#combine( #1( colon cancer ) )`. 

![Gender Distribution](image)

Figure 6.4: Gender Distribution

- Male: 40%
- Female: 40%
- Unknown: 18%
Concept Normalisation

Concept Normalisation takes a concept span and maps it to a CUI in the UMLS Metathesaurus. To perform this task on the TREC collection, we combine the span matching of the i2b2/VA MER system with the Concept Normalisation system described in Section 3.4. Processing is performed on both the query set and document collection.

To leverage Concept Normalisation in an IR context, we again turn to Indri’s field matching criteria. In the document collection, each sentence in the document has its concepts extracted and normalised. These normalised CUIs are then appended to the document text as CUI fields, e.g. \(<\text{CUI}>C0522224</\text{CUI}>\). By passing this system over the query set, we allow for matching based on normalised concepts, e.g. \(#\text{combine}(\text{children}\#1(\text{cerebral palsy})\ C0522224.\text{CUI})\).

6.4.4 Assertion Classification

The final aspect of including NLP information to the Clinical IR workflow is the inclusion of the assertion status of medical conditions. Assertion Classification is performed using the multi-class classification system discussed in Section 4.4.2.

The application of the system builds on the Concept Normalisation approach. First, assertion classification is performed on the span identified in the text. Once we have determined the assertion status of the span, we then apply the assertion status to the normalised CUI matching the text span. So the query \(#\text{combine}(\text{children}\#1(\text{cerebral palsy})\ C0522224.\text{CUI})\) becomes \(#\text{combine}(\text{children}\#(\text{cerebral palsy})\ C0522224.\text{PRESENT})\). The same process is performed on the document collection such that the query will match documents containing the field \(<\text{PRESENT}>C0522224</\text{PRESENT}>\). As this system extends from the i2b2/VA Assertion Classification task, there are six possible assertion fields for querying: \text{PRESENT, ABSENT, POSSIBLE, HYPOTHETICAL, CONDITIONAL and SOMEONE_ELSE}.

In the next section, we will describe the Clinical IR workflow and how the components described in this section combine to create a search system to investigate what ways NLP can aid the search process.
6.5 Clinical IR Workflow

So far in this chapter, we have discussed the individual tools and components that combine to form the Clinical IR system. At this point we will describe how these components are combined to form a Clinical Information Retrieval workflow. Figure 6.5 shows the overall workflow of the system. For both queries and the document collection the workflow remains largely the same, only the format of the outputs is changed.

![Clinical IR system Workflow](image)

Figure 6.5: Clinical IR system Workflow

The first stage in processing a text is the annotation of demographic information. In the document collection, age and gender information is marked by XML mark-up. In a querying context, age and gender information is added using the #scoreif operator, as shown in the query below.

- \#scoreif( child.agegroup #combine( admitted #1( cerebral palsy ) ) )

The #scoreif operator requires that only documents with an agegroup value of child will be returned. The documents which contain this field and value pair are then ranked according to the query given by the #combine statement. Initial experiments in which the age field is treated as a term, had little impact on results. Therefore, the more strict #scoreif operator is utilised to ensure only the required demographic is returned.

The second stage performs Medical Entity Recognition on the input text. This may have two possible outputs. If we wish to use field-matching, the type
of concept is added as a field to the entity span. However, if the IR system requires only to annotate phrasal boundaries, this MER tool annotates the phrasal span in the queries. If we are processing queries, the extracted span of text is passed to the query expansion module. Expansion may then be performed using either the MeSH or UTS expander systems.

The next stage in the process normalises the spans extracted by the I2B2 MER system to CUIs in the UMLS using the Concept Normalisation system described in Chapter 3. Following the annotation of CUI information in the text, the assertion status of extracted medical conditions may be classified using the multi-class classification system discussed in Section 4.4.2.

The modular nature of the proposed clinical Information Retrieval system allows various components to be switched on and off. This functionality allows for the identification of the most contributory aspects of NLP to the Information Retrieval process.

From the user’s perspective, these systems combine behind the scenes in that the user is still only required to enter their natural language query. No knowledge of the structured query language is required of the user. Instead, the tools described in this chapter add in the required structural operators while performing stopword removal and re-structuring of queries to leverage information in the document collection.

6.6 Summary

In this chapter, we presented an overview of the experimental setting and workflow of a Clinical Information Retrieval system in the investigation of this thesis. In particular we have discussed the document collection and query sets that are to be used in evaluating the system, as well as a discussion of the system’s constituent components and how NLP can be integrated into the IR workflow.

The data and query sets used in evaluating the Clinical Information Retrieval System

To evaluate the Clinical Information Retrieval system, queries, datasets and relevance judgements from the 2011 and 2012 TREC Medical tracks. TREC Medical track provides a standardised forum for the comparative evaluation of Information Retrieval systems. In using these datasets and relevance judge-
ments, work performed in this thesis may be compared to further work in the domain.

The functionality and features of the Indri Retrieval system.

The Indri retrieval system forms the basis of the Clinical IR system investigated in this thesis. In this chapter, we discussed the aspects of its query language that facilitate the inclusion of Natural Language Processing systems into the Information Retrieval workflow.

The development of tools used in automatically translating natural language queries into Indri’s structured query language.

In this chapter, we have provided a discussion of tools and systems that allow the translation of natural language queries into structured queries that leverage Indri’s structured query language and the output of natural language processing systems. The tools integrate demographic information, phrasal matching, field-based matching, concept normalisation and assertion classification into queries. Further to this, we have developed methods to allow automatic query expansion using both MeSH and UMLS ontologies. These tools remove the need of a user having knowledge of Indri’s complex query language. Instead, these tools combine to translate a natural language query into Indri’s structured query language. In doing so, we gain all the advantages of structured retrieval while avoiding costs of training healthcare professionals in a complex querying language.

The integration of NLP techniques with the Indri Retrieval system.

To integrate NLP techniques with the Indri Retrieval system, we leverage its structured query language. Demographic information such as age and gender are added to documents as XML marked fields. This information, along with the #scoreif operator ensure that patients of the correct demographic are returned and then ranked according to medical inclusion criteria. Field matching is further leveraged with concept normalisation and assertion classification methods. A Medical Entity Recognition system provides many services to the Information Retrieval workflow. Firstly, it may be used directly in annotating fields for medical conditions, treatments and tests. A more generalisable approach simply uses the system to annotate phrasal spans in queries. A final contribution of the MER system, is its extraction of spans may be then passed to expansion systems for accurate query expansion.

In the next chapter, we will provide and discuss experimental results in applying NLP to the IR workflow.
CHAPTER
SEVEN

EXPERIMENTAL EVALUATION OF A STRUCTURED CLINICAL INFORMATION RETRIEVAL APPROACH

7.1 Introduction

In this chapter, we provide an experimental evaluation of the Clinical Information Retrieval system presented in Chapter 6. The Clinical IR system leverages structural information in both the queries and document collection. Structural information is added to queries and documents through the application of NLP systems. Namely, we investigate the use of Medical Entity Recognition, Assertion Classification and Concept Normalisation in the context of an Information Retrieval setting. All systems applied to the task have been evaluated in standalone testing either as part of shared tasks, as in the CLEF MER system, or using previous shared task datasets available to the community, such as the i2b2/VA dataset. To evaluate the IR approach we use TREC queries and datasets from the 2011 and 2012 Medical Tracks, both available to the community. In using these standardised and available datasets, we aim to provide a basis for comparable evaluation of the combination of natural language processing and information retrieval techniques. Our experimentation in this chapter investigates the following research questions:

- What are the specific issues in applying MER to queries, rather than sentences?
- What are the effects of applying Medical Entity Recognition to an Information
Retrieval workflow?

- What are the effects of applying Assertion Classification to an Information Retrieval workflow?

- What are the effects of applying Concept Normalisation to an Information Retrieval workflow?

- What is the role of query expansion?

- Where does structural information aid searching, and what are its pitfalls?

- Do NLP tasks such as MER and Assertion Classification really contribute to the Information Retrieval task?

The structure of this chapter is as follows: In Section 7.2 we provide the experimental setting, recapping the metrics and query sets used in evaluating the performance of the system; Section 7.3 evaluates the performance of annotating medical entities in the query set; Section 7.4 describes an experimental evaluation of the NLP and IR hybrid system on the TREC Medical track 2011 and 2012 query sets.

7.2 Experimental Setting

In this section, we will define the exact setting used for the evaluation of the Information Retrieval systems put forward in this chapter. First, we will review the datasets used in evaluating the systems, followed by a definition of the evaluation methodology.

7.2.1 Experimental Datasets

Our experimental evaluation in this chapter falls under two evaluation sets. Both are based on the TREC Medical Track. The first on the 2011 query sets, with the second evaluation on the 2012 queries.

The dataset for the TREC Medical track comprises of 101,711 electronic health records extracted from BLULab’s NLP repository. For the retrieval task, the unit of retrieval is a visit. A visit contains one or more records, which relate to a given patient’s visit to a health facility. The document collection for the TREC task is composed of 17,266 visits. The reports can range from progress
notes and discharge summaries to radiology and surgical pathology reports. The query sets are taken from both the 2011 and 2012 tasks. The 2011 track contains 39 queries, and the 2012 track contains a further 49 natural language requests.

### 7.2.2 Experimental Evaluation

In Section 5.2.2, we introduced several metrics in the evaluation of Information Retrieval techniques. For our experiments, our primary metric is the BPref metric, defined as follows:

\[
BPref = \frac{1}{N} \sum_{r} 1 - \frac{|n \text{ ranked higher than } r|}{N}
\]  

(7.1)

This formulation allows the metric to be robust in cases where relevance judgements are incomplete.

### 7.3 Applying Sentence-based MER to Queries

Medical Entity Recognition (MER) in a shared task setting requires the recognition of medical entities that occur in medical health records. As described in Chapter 3, this task is performed on a sentence by sentence basis. For each token \( t \) in a sentence \( S \), the MER aims to assign a label to \( t \) to determine if it is the member of a medical entity. The sentence is a cohesive structure, every token has a role to play and each token’s placement in a sentence can be the driving force in conveying the overall meaning behind its construction. Information Retrieval queries do not have this same structure. In a typical IR setting, keyword queries are a set of interchangeable tokens. However, this is not always the case. As discussed in Chapter 5, queries may feature term dependencies or may be expressed as a request in a grammatical string e.g. *Patients with diabetes*. For this task, we wish to extract medical entities in order to discover term dependencies so that more accurate automatic querying may be performed. In our first set of experiments, we investigate the performance of the MER system developed on the i2b2/VA dataset as described in Section 3.3.2.

**Applying MER to Natural Language Queries**

The MER system used to annotate the TREC query sets is the same as that
described in Section 3.3.2. However, its model is built using both training and test documents from the i2b2/VA subset. The system is evaluated on both TREC Medical track 2011 and 2012 query sets. They contain 35 and 50 queries, respectively. The task is defined to be the same as that of the i2b2/VA challenge in that the system must correctly identify Problem, Treatment and Test spans in the query. Gold standard annotations were generated by the author. Table 7.1 shows the results of evaluating the system on the TREC query sets in reference to performance on the i2b2/VA test set performance.

<table>
<thead>
<tr>
<th></th>
<th>Exact</th>
<th>Inexact</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>i2b2 Test Set</td>
<td>0.77</td>
<td>0.74</td>
</tr>
<tr>
<td>TREC 2011 Queries</td>
<td>0.65</td>
<td>0.68</td>
</tr>
<tr>
<td>TREC 2012 Queries</td>
<td>0.76</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Table 7.1: Applying MER to TREC 2011 and 2012 Query Sets: Evaluation

For the exact matching criteria, performance dropped on the query set evaluations while improvements were seen on the inexact metrics. There are several reasons for this, that also highlights issues in MER for queries. The first problem area is disjunction. For example, given Query 1 below, the disjunctive span, *hip or knee surgery*, is incorrectly annotated, shown in Query 2. This is a difficult instance to annotate given a traditional entity recognition approach, as the text on its own cannot be divided as it is, given that surgery may be modified by either *hip* or *knee*. One approach could be to expand the span into two separate phrases e.g. `#1(hip surgery).treatment or #1(knee surgery).treatment` however this causes further difficulty in requiring a module to discover and reason about disjunctive statements. These difficulties further motivate the task of Concept Normalisation which may map the span to concepts representing the two forms of surgery.

1. Patients admitted for *hip or knee surgery*

2. Patients admitted for `#1(hip).treatment or #1(knee surgery).treatment`

The second issue for the system in tagging queries is shown in Query 3 below where both abbreviations and their expansions are given in sequence. This leads to the common error where they are tagged as a single entity e.g. `#test(positron emission tomography PET)`. This is a factor in higher performance on inexact metrics as the definition of the inexact metrics, i.e. overlapping criteria, counts two true positives as the identified span overlaps with two entities in the gold standard, i.e. *positron emission tomography* and *PET*.
3. Patients who had positron emission tomography (PET) or computed tomography (CT).

These occurrences are typical of keyword queries, where the user further clarifies terms. Again, a Concept Normalisation approach may be used to alleviate such difficulties.

In summary, the MER approach works best on the TREC query set which uses natural language requests rather than keyword based querying. However, there are issues with disjunction and keyword based information that may be addressed through the usage of concept normalisation.

7.4 Experimentation on the TREC Medical Track 2011/2012 Dataset

The TREC Medical track provided a forum for the evaluation of Information Retrieval techniques on collections of electronic health records. The task itself is defined to reflect the real world scenario where a healthcare professional may use a collection of clinical records to identify cohorts who are eligible for a clinical trial. For this search task, patients are represented by a visit document, as described in Section 6.2. The inclusion criteria for the trial are expressed in a natural language request, e.g. Female patients admitted with migraines. In this section, we investigate the application of NLP tools described in Chapter 3 and Chapter 4 to the TREC retrieval task.

7.4.1 Analysing the Impact of Annotating Problems, Treatments and Tests

Our first experiments analyse the use of a Medical Entity Recognition system, based on the i2b2/VA concept identification task, to the retrieval of clinical records. Two approaches are taken in this experimentation. The first approach applies structure to both the collection set and the query set. This approach allows for field matching between queries and documents. Annotated spans are identified using problem, treatment and test tags in an XML-like mark-up schema. The second approach does not perform field matching, rather it simply identifies phrasal spans. The use of this methodology allows for the occurrence of term dependencies e.g. colon cancer.
For our experimentation, we evaluate on both TREC 2011 and 2012 query sets. We focus primarily on the BPref metric as it was the primary metric available at both tasks. As a reference point, we also include a keyword baseline run that uses Indri’s retrieval model. The initial runs for our i2b2/VA based experiments are defined as follows:

- **Sys_MER**: Uses field-matching to identify concepts in the text. There are three possible field types, namely: Problem, Treatment and Test. Fields are identified in queries and the corpus by the MER system described in Section 3.3.2.

- **Sys_Phrasal_1**: Uses phrasal matching in queries. Query boundaries are identified in queries and the corpus by the MER system. An initial window of size 1 is used for phrases.

- **Gold_MER**: Uses field-matching to identify concepts in the text. Fields are identified in queries by a human annotator and in the corpus by the MER system described in Section 3.3.2.

- **Gold_Phrasal_1**: Uses phrasal matching in queries. Phrasal boundaries are identified in queries by a human annotator and in the corpus by the MER system. An initial window of size 1 is used for phrases.

All of the above systems are expanded using related MeSH terms as described in Section 6.4.1. The performance of these initial runs with respect to a keyword baseline is shown in Table 7.2. These experiments aim to allow us analyse any effects of MER on retrieval performance with respect to manual annotations. The MER system used in annotating spans recognises only continuous sequences, rather than disjoint sequences. This system was chosen as none of the queries contained disjoint concepts.

<table>
<thead>
<tr>
<th>System</th>
<th>TREC2011</th>
<th>TREC2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyword</td>
<td>0.456‡</td>
<td>0.324‡</td>
</tr>
<tr>
<td>Gold_Phrasal_1</td>
<td>0.333‡</td>
<td>0.318‡</td>
</tr>
<tr>
<td>Sys_Phrasal_1</td>
<td>0.319‡</td>
<td>0.284‡</td>
</tr>
<tr>
<td>Gold_MER</td>
<td>0.118</td>
<td>0.093</td>
</tr>
<tr>
<td>Sys_MER</td>
<td>0.115</td>
<td>0.087</td>
</tr>
</tbody>
</table>

Table 7.2: BPref scores of IR systems leveraging MER on TREC 2011 & 2012 query sets. ‡ indicates the best run by a statistically significant margin. † indicates a statistically significant difference in performance than the next ranked run.
The keyword based system is the top performing run by a statistically significant margin. Ultimately, the strict nature of field matching whereby we aim to map exact spans between query and document had a negative impact on performance, for both the human annotated and machine-learning system annotated queries (i.e., Gold_MER, Sys_MER). This is further compounded by the system’s imperfect MER performance. For example, given a document that contains the concept span diabetes, if it is not recognised by the system, in both query and the document, it will create a knock-on error where the document is deemed irrelevant to the query, as fields are incorrectly annotated.

The phrasal system relaxes this requirement as it only accounts for terms occurring in order, rather than an exact field. As a result, performance improves for the phrasal systems, with the gold standard queries slightly outperforming the machine annotated queries. Though all systems are outperformed by the keyword approach, a cursory analysis shows that the phrasal matching techniques outperform the keyword approach on more complex queries.

To investigate this further, we relax the phrasal matching criteria to an unordered window of three. This window was chosen as an optimal window size from several sizes in testing. We then combine keyword and phrasal approaches to analyse if the deficiencies of one approach can be addressed by the superiority of the other. Rankings are combined through the ComboMNZ method, whereby for each document retrieved by the systems, the average of their scores is used as a new ranking score. This creates four new runs for analysis:

- keyword_gold_phrasal_1 combines keyword retrieval with Gold_Phrasal_1.
- keyword_gold_phrasal_3 combines keyword retrieval with Gold_Phrasal_1 where the window size has been increased to 3.
- keyword_sys_phrasal_3 combines keyword retrieval with Sys_Phrasal_1.
- keyword_sys_phrasal_3 combines keyword retrieval with Sys_Phrasal_1 where the window size has been increased to 3.

1Significance is calculated using the Paired Student's t-test with a p-value = 0.05, for all experiments in this chapter.
The performance of the new combined runs are presented in Table 7.3 below.

<table>
<thead>
<tr>
<th>System</th>
<th>TREC2011</th>
<th>TREC2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>keyword_gold_phrasal_3</td>
<td>0.388†‡</td>
<td>0.345†‡</td>
</tr>
<tr>
<td>keyword_gold_phrasal_1</td>
<td>0.381‡</td>
<td>0.336‡</td>
</tr>
<tr>
<td>keyword_sys_phrasal_1</td>
<td>0.357</td>
<td>0.332‡</td>
</tr>
<tr>
<td>keyword_sys_phrasal_3</td>
<td>0.359</td>
<td>0.315</td>
</tr>
</tbody>
</table>

Table 7.3: BPref scores of combined IR systems using MER on TREC 2011 & 2012 query sets with Relaxed Windows. † indicates the best run by a statistically significant margin. ‡ indicates a statistically significant difference in performance than the next ranked run.

The combination of phrasal approaches with the keyword baseline improved the performance of the system over the phrasal systems on their own. However, the performance is still short of the Keyword baseline. Despite this, there are some identifiable traits. Firstly, we see that gold standard queries outperform the system generated queries. This shows that errors in the recognition of entities propagate throughout the workflow. From their identification to their expansion and finally in the retrieval stage. Secondly, we see that an increase in window size is met with an improvement in performance. While dependency information is important to retrieval, ensuring all items are ordered in a span of one token is too strict of a criterion as it does not account for any linguistic variation in the text, e.g. right arm fracture vs. fracture of right arm.

**What are the key findings from these experiments?**

Though the keyword based approach achieved the best performance across TREC 2011 and 2012 query sets, there are key areas where a phrasal based approach outperforms the baseline. For example, the following queries saw a phrasal approach outperform the baseline:

- **Query 153**: Patients admitted to the hospital with end-stage chronic disease who are offered hospice care. Keyword approaches face difficulty with verb-ose queries featuring several multi-term concepts.

- **Query 156**: Patients with Diabetes exhibiting good Hemoglobin A1c Control (<8.0%). While a keyword approach may identify the concept diabetes correctly in a corpus, the annotation of the concept as diabetes exhibiting helped correctly identify reports where patients with diabetes exhibit a trait.

- **Query 165**: Patients who have gluten intolerance or celiac disease. Keywords
take no account of term dependencies. General terms such as intolerance and disease severely hampered performance.

In short, the keyword approach outperformed phrasal methods where the query was simple, containing few concepts terms that consist of single tokens. Conversely, phrasal matching worked best with concept dense queries. To improve performance, both approaches are combined with an increase in window size.

Figure 7.1 shows the performance of a combined system using a window of size 3 with respect to keyword and phrasal approaches with a window of size 1. While the combined system improved performance of the phrasal approaches, it was still outperformed by the baseline on the TREC 2011 query set. The combined system significantly improves performance on verbose queries e.g. query 106, Patients who had positron emission tomography (PET), magnetic resonance imaging (MRI), or computed tomography (CT) over previously restricted phrasal windows.

Performance of combined approaches was much more positive on the TREC 2012 query set, as shown in Figure 7.2. The Gold_Phrasal_1 run outperformed all other runs notably on queries 151, 153, 176, 183. All of these queries contain multi-term concepts, where one of the terms is highly gen-
eralisable, e.g. disease in liver disease (Query 151) or acute in acute vision loss (Query 183). The combination phrasal information with keyword approaches improved keyword performance where queries contained several multi-term concepts, as in Queries 145 and 176 below.

Figure 7.2: MER and IR on TREC 2012 Query set : BPref breakdown

- **Query 145**: Patients with lupus nephritis and thrombotic thrombocytopenic purpura

- **Query 176**: Patients with Heart Failure (HF) on Angiotensin-Converting Enzyme (ACE) Inhibitor or Angiotensin Rec.

Key issues remain with query expansion such as Query 179, Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression, where the general concept term atypical antipsychotics featured no expansions. Atypical antipsychotics is a more unlikely term than specific atypical antipsychotic drug names e.g. Clozaril, Risperdal, Zyprexa, that may be found through query expansion methods.
7.4.2 Concept Normalisation and Query Expansion Coverage

In our experimentation using structured retrieval techniques, it was shown that linguistic variation, an inadequate vocabulary for expansion and overly strict phrasal bounds hampered its performance. In this section, we address these issues through the application of Concept Normalisation and an expansion technique using a larger volume of data. For Concept Normalisation, the system developed as part of the CLEF/ShARe task as described in Section 3.4 is applied to the query set and document collection. Query expansion is investigated using the UTS Expander system which was fully described in Section 6.4.1. The experiments in this section also leverage the UMLS Metathesaurus rather than MeSH headings in order to take advantage of the expansive vocabulary and relationships it contains.

First, we explore the task of query expansion. Query expansion requires a delicate balance. We require an expansion of concept terms so that we may encapsulate terminological variation, however if an expansion is too vast we may see the introduction of noisy terms to the query that may harm overall performance. Table 7.4 demonstrates the effect of the UTS and MeSH expansion techniques on the number of terms within a query.

<table>
<thead>
<tr>
<th>Query Set</th>
<th>TREC 2011</th>
<th>TREC 2012</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>308</td>
<td>347</td>
<td>7.7</td>
</tr>
<tr>
<td>MeSH</td>
<td>564</td>
<td>722</td>
<td>15.1</td>
</tr>
<tr>
<td>UTS</td>
<td>5882</td>
<td>7200</td>
<td>153.9</td>
</tr>
</tbody>
</table>

Table 7.4: Effects of Expansion Resource Choice on Query Length

Though the use of MeSH based expansion doubled the number of terms in a query, it was often the case that concept terms, e.g. hearing loss had no expansions. At the other end of the spectrum, the UTS expansion system increased the number of terms in a query by 20 times the original query length. The original UTS expansion method first retrieved 10 CUIs for a given phrase, and allowed up to ten related terms as expansions. Furthermore, it allowed for relations other than synonymy. To further refine this method, and preempt any noise introduced through the UTS, we will also look at allowing only synonymous relations as expansions and choosing concepts that are best matches according to the UTS. Conversely, given the relatively well performing baseline, we investigate the combination of keyword information with the UTS expansion method in order to further discover pertinent terms that may
be hampered by unnecessary dependencies. Therefore we set out four runs to investigate the effects of query expansion. These runs are defined as follows:

- **MeSH** uses the MeSH expansion strategy described in Section 6.4.1.
- **UTS** uses the UTS Expander system described in Section 6.4.1. The system identifies all matching concepts in the UMLS Metathesaurus and expands on all relations that are of the the types **RL** (Synonymous), **RN** (Narrower), **RB** (Broader) or **RO** (Relation other than **RN**, **RL**, **RB**) as specified in Table 6.1.
- **UTS_Ref** uses the UTS Expander system to identify concepts that attain a maximum matching score, with only synonymous (**RL**) or narrower relations (**RN**) used to expand terms.
- **UTS_Key** uses **UTS_Ref**’s expansion method. The expansion terms are close to the original term and addresses its restrictive nature by including keywords as well as phrases generated from these expansions, e.g. if we discover the expansion **dental caries**, we create three additional expansion terms for the query. Namely “dental”, “caries” and the phrase itself “dental caries”.

These four runs are designed to investigate the effects to UMLS and MeSH based expansions, as well as the effects of more refined approaches of seeking expansion terms in ontologies. All of the runs use phrasal matching as generated through the i2b2/VA system as described in the previous section. Terms within an identified concept must occur within tokens of one another in a document, regardless of order. The resulting performance of the expansion methodologies on the TREC 2011 and 2012 query sets are shown in Figures 7.3 and 7.4, respectively.

<table>
<thead>
<tr>
<th>System Name</th>
<th>TREC 2011</th>
<th>TREC 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTS_Key</td>
<td>0.397</td>
<td>0.294†‡</td>
</tr>
<tr>
<td>MeSH</td>
<td>0.319</td>
<td>0.278‡</td>
</tr>
<tr>
<td>UTS_Ref</td>
<td>0.395‡</td>
<td>0.236‡</td>
</tr>
<tr>
<td>UTS</td>
<td>0.349‡</td>
<td>0.227</td>
</tr>
</tbody>
</table>

Table 7.5: BPref Performance of Expansion Strategies on TREC 2011 & 2012 query sets. † indicates the best run by a statistically significant margin. ‡ indicates a statistically significant difference in performance than the next ranked run.
The performance of each system fluctuates on each query, highlighting both the differences in query types as well as the weaknesses and strengths of each expansion method. The shortcomings of MeSH were most noticeable for queries 101, 123, 124 and 133. Query 101 specified *Patients with hearing loss* where the MeSH system retrieved no expansions, thus missing key terms such as *deafness* or *ear damage*. While expansions were found for query 123, they were largely inadequate. The only expansions generated for *chest pain* were *chest* and *pain* using the MeSH system. Where the UTS system found expansions such as *chest wall tenderness*. For query 124, MeSH introduced more noisy terms (*neoplasms*, *metastasis*). This is in spite of it being the most tentative expansion approach in terms of how far reaching its expansions are. The MeSH based system found no expansions for *herbal products* in query 133, while UTS approaches expanded to the more common *herbal medicines*. However, this limited expansion also worked to its advantage in cases such as query 113, where the simple request for *colonoscopy adenocarcinoma* bested the more expansive approaches.

The varying application of UTS expansions also had a clear effect on performance. As shown in Table 7.5, the hugely expansive UTS run was significantly outperformed by all other runs. The most notable exceptions are queries 101, 109 and 132. For query 132, the UTS based expansions for more general terms
in relation to surgery such as operating room and procedure which create an overall sense for the query. In query 101, MeSH created no expansions for hearing loss. While the UTS system expanded to over twenty terms, it gained crucial yet simple expansions such as deafness and impairment. UTS_key generated keywords from the UTS expansions. Its more generalisable performance can be noted across several queries. For example, in query 119 included expansion terms such as metabolic and endocrine which are not mentioned explicitly in the query, but feature in top ranking documents.

Figure 7.4: Topic by topic: Query Expansion on TREC 2012

On the TREC 2012 query set, there are more noticeable differences in performance, owing to the more varied style of queries. The UTS and UTS_Ref runs are significantly outperformed by MeSH and UTS_key systems. MeSH outperforms on queries which are relatively simple e.g. query 158, Patients with adult respiratory distress syndrome. The inadequacy of MeSH expansions were shown in queries 148 and 165. For example, for query 148, Patients acutely treated for migraine in the emergency department MeSH generated no expansions. Through missing even simple expansions, such as headache for migraine the MeSH system misses out on opportunities for near optimal performance. The simple expansion of headache for migraine by the UTS_key system led to most top ranking visits being retrieved. This improvement in performance also highlights the strength of the UTS_key approach. The first expansion returned by
the UTS is severe headache. By splitting this into keywords as well as a phrase it gave the query more generalising power in retrieving documents with discussions of headaches along with mentions of severity.

From our experimentation, we can see the impact of different query expansion techniques can often depend on terms in the queries. For example, while the UTS expander aided the task in identifying expansions for more nuanced phrases such as herbal products, the relative restrictiveness of the MeSH based system tended to prevent over-generalisations. We will now discuss another, more novel methodology in resolving mismatch issues between query and document.

**Concept Normalisation for Information Retrieval**

Concept Normalisation addresses vocabulary mismatch issues through mapping concepts in queries and documents to identifiers in an ontology. The effect of mismatch between terms is reduced as documents and queries are mapped to the same semantic space. In this section, we apply the concept normalisation system from Section 3.4.2 to queries and documents in the TREC collection to aid the search task. To perform an evaluation of the application of Concept Normalisation, we analyse two runs:

1. **C_Norm** applies the concept normalisation system to both the queries and documents. Querying is performed using phrasal matching and field matching between CUIs.

2. **C_NormUTS** applies concept normalisation in conjunction with expansion of concept terms using the UTS Expander system from run UTS_Key.

Figure 7.5 and Figure 7.6 show a breakdown of the performance of each run on the TREC 2011 and 2012 query sets, respectively. For reference purposes, the MeSH query expansion system performance is also included. The average performance of the three runs is provided in Table 7.6. Notably, **C_Norm** outperforms the other two systems by a statistically significant margin. The introduction of query expansion terms to the Concept Normalised run brought a significant drop in performance.

Throughout the queries in Figure 7.5 and Figure 7.6, the Concept Normalisation systems comprehensively outperform the standard phrasal matching and query expansion approach. The most notable improvements are seen in queries 104 and 107. Query 104 specifies the use of robotic surgery, however this is
Table 7.6: BPref Performance of Normalisation Strategies on TREC 2011 & 2012 query sets. † indicates the best run by a statistically significant margin. ‡ indicates a statistically significant difference in performance than the next ranked run.

<table>
<thead>
<tr>
<th>System Name</th>
<th>TREC 2011</th>
<th>TREC 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_Norm</td>
<td>0.436†‡</td>
<td>0.303†‡</td>
</tr>
<tr>
<td>MeSH</td>
<td>0.319</td>
<td>0.278‡</td>
</tr>
<tr>
<td>C_NormUTS</td>
<td>0.423‡</td>
<td>0.247</td>
</tr>
</tbody>
</table>

Figure 7.5: Topic by topic: Concept Normalisation on TREC 2011

not a very common term. As a result, ensuring that both terms occur in a span could indeed be harmful to performance. In using a normalisation approach we may map the concept to the CUI C0543467 (Operative Surgical Procedures) which is more generalisable among the corpus. For query 107, we have the request Patients with ductal carcinoma in situ (DCIS). There are many variations in abbreviating the term as well as available synonyms (UMLS lists 258 atoms). Normalising the concept again reduces the need in accounting for variation in the text, with a gain in performance. Notably, the introduction of expansions reduces overall performance. As normalisation already generalises terms, incorrect span matching and too general text expansions negate performance. For example, query 178, Patients with metastatic breast cancer, the normalisation system normalises to four concepts, all of which directly identify carcinoma in breasts. This normalisation achieves a BPref score of 0.759. However, the
expansion introduces considerable noise such as Disorder calcium NOS. As a result, we see a considerable drop in performance to 0.409.

![Figure 7.6: Topic by topic: Concept Normalisation on TREC 2012](image)

On the TREC 2012 query set, we see that expansions seem to harm the performance of runs that use concept normalisation. While there is positive performance from the normalisation approach, it is also outperformed by the expansion approach in certain instances. For example, query 161 seeks Patients with adult respiratory distress syndrome. However, the concept normalisation system incorrectly normalises to CUI’s for distress and syndrome rather than the correct respiratory distress syndrome, adult. Such instances highlight the leading issues of a wholly concept based retrieval approach in that if we totally rely on a mapping, it is very easy to propagate errors. In the next section, we shall refine our approach through the application of demographic information and the assertion classification of medical conditions.

### 7.4.3 Assertion Classification and Demographic Information

In this section, we describe methods of introducing information to querying that allows a more specific search. While the inclusion criteria for a clinical trial may require details on medical conditions, treatments or tests, it may also
require key demographic information such as the age or gender of a patient. The determination of gender and age in queries is performed by the rule based recognition system discussed in Section 6.4. Also in this section we introduce assertion classification to the querying of records. The machine learning based Assertion Classification system described in Section 4.4.2 is used to perform this task. Given the overly restrictive nature of field based matching as found in the experiments of Section 7.4.1, the system will assert on normalised concepts, CUIs as in Section 7.4.2’s experiments. In asserting on CUIs for retrieval it removes the strict matching on text criteria and allows querying to be more generalisable.

In the previous section, it was shown that concept normalisation can outperform methods which leverage query expansion. Concept normalisation allows for the generalisation of concepts, while minimising errors such as those caused by overly strict phrasal matches or the introduction of noisy results using keyword approaches. However, the use of concept normalisation creates a system which is precision focused and as a result its performance may be severely hampered by even a single incorrect normalisation. For these reasons, our system will now use a keyword based approach coupled with the concept normalisation system.

**Age and Gender Filtering**

Inclusion criteria for clinical trials may include a patient’s current health, such as medical conditions or treatments they have undergone. However, inclusion criteria may also specify demographic information such as age and gender. In this section, we examine the effects of applying gender and age filtering to queries and documents. For these experiments, we present three runs as follows:

- **keyword_C_Norm** consists of a keyword query, with concept normalisation information extracted from queries by a medical entity recognition system. This run aims to combine the effectiveness of concept normalisation on queries containing multi-word concepts, with the generalisation ability of keywords.

- **keyword_C_Norm_age** builds on the previous run with the inclusion of age filtering. According to this criteria, only documents within a specified age range may be returned. This run aims to determine the usefulness of exact age matching criteria.
• `keyword_C_Norm_age_gen` builds on the previous run through the inclusion of gender information. The same strict matching is implemented as the age run, as only documents containing the specified gender may be returned. This run is used to assess the impact of gender information in querying.

The performance of the runs with respect to the keyword baseline is given in Table 7.7.

<table>
<thead>
<tr>
<th>System Description</th>
<th>TREC 2011</th>
<th>TREC 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>keyword_C_Norm</td>
<td>0.505†</td>
<td>0.338†‡</td>
</tr>
<tr>
<td>keyword_C_Norm_age</td>
<td>0.500‡</td>
<td>0.310</td>
</tr>
<tr>
<td>keyword_C_Norm_age_gen</td>
<td>0.493‡</td>
<td>0.310</td>
</tr>
<tr>
<td>keyword</td>
<td>0.456</td>
<td>0.324‡</td>
</tr>
</tbody>
</table>

Table 7.7: Concept Normalisation Approaches with Age and Gender filtering. † indicates the best run by a statistically significant margin. ‡ indicates a statistically significant difference in performance than the next ranked run.

Notably, the combination of keywords and the concept normalisation approach **improved performance beyond all other runs**, outperforming the baseline considerably on the TREC 2011 dataset, with a slight difference in performance on TREC 2012 queries. The inclusion of age and gender filtering saw a drop in performance. A key cause of this drop in performance is the coverage of `gender` and `agegroup` tags. As discussed in Section 6.4, 2,500 visits feature a null gender tag, with 3,000 visits containing no known age information. With strict matching for age and gender criteria these documents are readily discounted from relevance judgements even though other content in the document may deem it relevant.

The difference in performance when implementing age and gender filtering seemingly does not justify its inclusion. However, it must be said that the query set does not require the information to attain top performance. For example, only two queries among the TREC 2011 and 2012 collection require gender information: Query 109 specifies osteopenia, with query 112 requiring breast cancer patients who have had mastectomies. Both illnesses have a much higher frequency among women, and as such the majority if not all top ranking documents will discuss female patients. This results in no impact where gender filtering is applied, or negative impact if relevant documents are missing gender information. While not applicable on these query sets, it is clear that in a real world setting dealing with many different conditions, with
more equal distributions, filtering would be required.

**Analysing the effects of Assertion Classification**

NLP tasks such as Assertion Classification are often cited as tools that may aid the Information Retrieval process. In these experiments, we analyse the impact of performing assertion classification on concepts in the document collection and query set.

The Assertion Classification used is the SVM based approach described in Section 4.4. It is applied to documents and queries in a three stage process. First, entities are detected in the text using the MER system as described in Section 7.3. The Assertion Classification system then classifies the assertion status of the identified entity. The Concept Normalisation system in Section 7.4.2 then maps the identified concept to a CUI in the UMLS. This CUI is then assigned the assertion status of the identified concept, i.e. *Present, Absent, Hypothetical, Conditional, Possible* and *someone else*. Combining this approach with Indri’s structured querying then retrieves documents based on assertion status.

Through the use of this approach, the system may ignore mentions of concepts that have an assertion status other than that denoted in the query. A breakdown of the system’s impact with respect to the top performing concept normalisation system is provided in Figure 7.7 and Figure 7.8 with a summary in Table 7.8.

<table>
<thead>
<tr>
<th>System Name</th>
<th>TREC 2011</th>
<th>TREC 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>keyword_C_Norm</td>
<td>0.505†‡</td>
<td>0.338†‡</td>
</tr>
<tr>
<td>keyword_C_Norm_Assert</td>
<td>0.405</td>
<td>0.280</td>
</tr>
</tbody>
</table>

Table 7.8: BPref results of Applying Assertion Classification to Concept Normalisation. † indicates the best run by a statistically significant margin. ‡ indicates a statistically significant difference in performance than the next ranked run.

On the 2011 TREC query set, the use of assertion classification was significantly outperformed by the *keyword_C_Norm* approach. A key aspect of this is the restrictiveness introduced by the approach as well as the fact that no queries had an assertion status other than positive. However, there are some indications among the queries where assertion classification aids performance. For example, given query 104, *Patients diagnosed with localised prostate cancer and treated with robotic surgery* *keyword_C_Norm* retrieved several records where the condition affected another family member. The use of assertion classification reduced the number of documents describing such conditions for this
The TREC 2012 query set showed the same trend where the restrictive nature of the assertion classification approach saw it outperformed by
keyword_C_Norm. But again, for certain queries it may improve performance where assertion status do not match such as in query 174. Such differences are slight, but provide an indication to the possible application of assertion classification in future work.

7.5 Summary

In this chapter, we detailed experiments in investigating the application of structural information to a clinical information retrieval workflow. Structural information was generated from the Medical Entity Recognition and Concept Normalisation systems described in Chapter 3 and an Assertion Classification system as evaluated in Chapter 4. We will now address our experimentation with respect to the research questions set out at the beginning of this chapter.

What are the specific issues in applying MER to queries, rather than sentences?

To analyse issues in Medical Entity Recognition for queries, a set of experiments were performed on the TREC 2011 and 2012 query sets. For the TREC query sets, Medical Entity Recognition performance largely mirrored that of its performance on the i2b2/VA test set, owing to the structured nature of TREC queries. However, issues are noted in cases of disjunction and the use of abbreviations in conjunction with their expanded form. In summary, the system is most effective where queries replicate sentence structure, with queries containing information in a keyword format proving the most difficult to annotate.

What are the effects of applying Medical Entity Recognition to an Information Retrieval workflow?

Initial attempts at introducing Medical Entity Recognition explored two avenues of application. The first, the development of a strict field-matching system. Field-matching requires that annotated concepts in a query must match an annotated span in the text. However, this field-matching technique was largely unsuccessful. This application of structure to the text had poor performance as its strict matching criteria fail to allow for any variation in the text. Given the high level of variation of expression in clinical reports, it is an issue that must be further explored should a field matching approach be adopted. Further, despite the MER system’s relatively successful performance, its generation of false negatives introduced further errors into the pipeline whereby missing annotations in documents lead to the IR system’s incorrect decision
of irrelevancy. Despite this exact matching failure, MER shows promise in the creation of phrasal queries where phrasal bounds are identified in the queries, thus accounting for term dependency information in the texts. While the system was initially outperformed by the keyword baseline, analysis has shown improvements on performance where queries contain several multi-word concepts.

What are the effects of applying Concept Normalisation to an Information Retrieval workflow?

Experiments in applying Concept Normalisation to clinical queries improve performance and provide the best performance among approaches discussed in this thesis. The use of Concept Normalisation addresses two issues. First, it reduces over generalisation of query terms found in texts, with no term dependency information included. A more common solution to this is the application of phrasal boundaries. However as discussed above, phrasal boundaries can often be too restrictive. Concept Normalisation addresses both these issues as no exact span is required, allowing generalisations while they must match in their semantic content overcoming issues of keyword approaches. It is a high precision method, so its application to keyword queries marries its precision with the recall ability of keywords. As a result this combination we can identify our best performing approach as combining concept normalisation methods with the standard keyword approach.

What are the effects of applying Assertion Classification to an Information Retrieval workflow?

Given the poor performance of the field matching approach, assertion classification is applied to the concept normalisation workflow. In doing so, we keep the top performance of the concept normalisation approach while adding a new facet of specificity to the searching criteria. As the query set contained only medical conditions with a present assertion status, its effects on information retrieval were limited. Despite this, instances of improved performance have been found in the results, primarily with the indication of medical conditions afflicting another individual.

What is the role of query expansion?

In this thesis two ontologies, MeSH and the UMLS Metathesaurus, were investigated to perform query expansion. While MeSH provided good quality expansions, there were far too many instances where expansions were not available for terms in the query. The expansiveness of the UMLS Metathesaurus alleviated this issue, though it was found to expand to multiple
noisy terms. A restriction was put in place to ensure that only relations of synonymy and of a narrower semantic type were put in place, which in turn saw an improvement in performance. Query expansion, in cases of no concept normalisation, are an integral part of retrieval. This is not only to address issues of vocabulary mismatch, but the use of UMLS expansions often feature terms that create a deeper semantic understanding of the text, e.g. discussion of anion gap acidosis would also feature metabolic rates.

**What is the impact of Age & Gender filtering?**

Demographic information such as the age and gender of a patient may be required inclusion criteria for a clinical trial. Given this, it is a logical step that such information should be included in the querying process for more accurate results. In our application of gender and age information, a set of pattern extraction rules are passed over the collection and queries to extract the required information. However, given that some documents do not contain age/gender information, its effects are limited. However in a full scale system, and in ensuring accurate results in the TREC collection it is an important step.

In our investigation of the application of Medical Entity Recognition and Concept Normalisation systems we have discovered that while MER systems are useful for identifying phrasal bounds and allow for functional expansion, *concept normalisation of extracted bounds are the most successful approaches in clinical information retrieval* and warrant further research in the domain in combination with text based methods.

In the next chapter, we will discuss the contributions of the experimental findings in this chapter, and our experiments in Chapter 3 and Chapter 4.
CONCLUSIONS AND FUTURE CONSIDERATIONS

In this chapter we discuss the conclusions and contributions of this thesis with respect to our initial research goals.

8.1 Introduction

This thesis investigates the applicability of recently proposed Clinical NLP techniques, to the task of Clinical Information Retrieval. To address this question, the thesis is divided into two research areas. The first area focusses on the development and evaluation of Clinical NLP methodologies, while the second aspect investigates the impact of their application to the Clinical Information Retrieval domain.

For our Clinical NLP investigation, three methods were discussed in the context of shared tasks in the domain. The first technique discussed is Medical Entity Recognition (MER) which aims to detect medical concepts such as disorders, treatments and tests in clinical texts. The system investigated leveraged lexical, syntactic and semantic information in a machine learning framework. Performance was moderate on i2b2/VA evaluation data, with more competitive results on CLEF/ShARe datasets. The second technique investigated the task of Concept Normalisation, where identified concepts from the MER task are mapped to unique identifiers in an ontology. The third task, Assertion Classification, determines the assertion status of identified medical disorders. The assertion status of a medical condition refers to whether or not a given medical condition is currently affecting the patient, e.g. given no edema the disorder edema has an assertion status absent. This system achieved results com-
petitive with the state-of-the-art and significantly outperforms manual rule-based approaches.

In our Clinical Information Retrieval experimentation, we applied Clinical NLP approaches to the retrieval task. The use of MER facilitated automatic phrasal querying, whereby term dependencies are taken into account for the retrieval process. Concept Normalisation also facilitates term dependency information through mapping both document and query concept text to unique identifiers in an ontology. Further to this it alleviates the issue of *vocabulary mismatch* whereby the concepts in text and query match, however documents have a low or zero rank due to the fact that lexically different terms are used to describe the same concept. Assertion Classification may be used to alleviate issues where there is assertion mismatch, e.g. the query may be searching for a patient with a specific disease but documents that refer to family member with this disease are retrieved.

### 8.2 Thesis Contributions

In this section, we detail the contributions of this thesis with respect to our research aims set out in Chapter 1. Similar to our conclusions, we subdivide the contributions into the Natural Language Processing domain in Section 8.2.1 and Information Retrieval in Section 8.2.2.

#### 8.2.1 Contributions to Natural Language Processing in the Clinical Domain

In this section, we analyse the contributions of this thesis to the Clinical NLP community. In particular, Section 8.2.1.1 discusses its contributions to Medical Entity Recognition and Concept Normalisation. Section 8.2.1.2 presents findings in Assertion Classification.

#### 8.2.1.1 Medical Entity Recognition & Concept Normalisation

- *What impact do different tagging formats have on MER performance?*
  
  **Finding:** Experimentation shows that for performing Medical Entity Recognition on the i2b2/VA challenge corpus, the BIESO tagging structure has shown higher rates of recall than the traditional BIO format. This
can be attributed to the BIESO tagging format’s ability to separate single-token concepts and concepts comprised of an above average number of tokens. Given the large variation in the textual representation of entities in the clinical domain, a BIESO based tagging format can improve recall.

- **What features contribute most to the MER task?**
  **Finding:** In our feature analysis of the Medical Entity Recognition system, we have discovered that token features, which analyse tokens preceding and succeeding an instance, contribute the most to the task. Affix, e.g. -tomy is a suffix of size four for the token tracheotomy, and semantic information allow the system to generalise its observations. For example, suffixes such as -tomy generalise to medical treatments while the suffix -aphy generalises to tests, e.g. angiography. The introduction of clustering information, along with part-of-speech information has been shown to improve the system’s precision as well.

- **How may a traditional MER system be adapted for the recognition of disjoint entities?**
  **Finding:** To extend our traditional MER system to the recognition of disjoint entities, a new tagging format was developed as part of this thesis, BIESTO, that tagged tokens between linked concept terms. The most notable finding in this experimentation highlighted the importance of semantic features which allowed the linking of disjoint terms through shared semantic characteristics. Despite this finding, performance on recognising disjoint spans is far from optimal and requires a new set of approaches to be fully resolved.

### 8.2.1.2 Assertion Classification

- **How do rule-based systems compare to machine learning performance for this task?**
  **Finding:** In our experimentation it was shown that machine learning approaches significantly outperform the rule-based ConText approach in the Temporal Grounding task. However, an interesting point to note is that for the Condition Attribution task whereby we require knowledge of the medical problem affecting the patient, or another individual in the report text, a simpler rule based system may suffice. This is based on the evidence that 99% of medical problems affecting another individual
occur under the *Family History* section heading.

- **What features give best-performance for Assertion Classification?**
  
  **Finding:** For the binary classification task of Temporal Grounding, our novel score-based features contributed most to the task. For Condition Attribution, given that 99% of medical problems affecting another individual occur under the *Family History* section heading, the heading under which a condition falls is the most prominent feature. In the case of the datasets used, any report type that discusses another individual, e.g. History & Physical examination, Discharge Summary will contain a heading of *Family History*. Reports that do not include this heading are ones that will not describe another individual, such as a radiology report. In extending the approach to a multi-class classification problem, we observe the importance of tokens surrounding the inspected medical problem. Further to this, heading information, the use of pattern matching as well as score-based features all contribute to the performance of the Assertion Classification system.

- **What machine-learning algorithm performs best at Assertion Classification?**
  
  **Finding:** Experiments in Temporal Grounding showed that the Random Forest algorithm provides top performance for the binary classification task. However, it was shown that it was inadequate for the multi-class classification problem. Instead the Support Vector Machine was chosen for the multi-class task. A One-Versus-One approach improved recall on minority classes, with a finely tuned RBF kernel achieving optimal performance.

### 8.2.2 Contributions to Information Retrieval in the Clinical Domain

In this section, we present and discuss the contributions of our findings of Information Retrieval experimentation in this thesis.

- **With shared tasks in Clinical NLP citing Information Retrieval as an application area, what NLP tasks contribute most to Information Retrieval?**
  
  **Finding:** In this thesis, we have investigated the use of Medical Entity Recognition, Assertion Classification and Concept Normalisation. Our experimental findings show that MER may be used to perform
phrasal querying, leading to improvements in performance where queries contain multiple term dependencies. Most notably, Concept Normalisation is the most effective NLP method for Clinical Information Retrieval, significantly improving performance over all systems developed in this thesis. Assertion Classification proved detrimental to the system’s overall performance.

- **What are the challenges faced in using MER to automatically translate natural language queries to a structured query language?**

  **Finding:** The application of MER to natural language queries largely reflected performance of the system on clinical reports. However, it is notable that constructs that are common to querying, such as disjunction e.g. “hip or knee surgery” pose the most difficulty to the MER system. This provides an indicator that while MER performs well in identifying concepts in queries, post processing rules may overcome issues with disjunction, or may be addressed by a concept normalisation approach.

- **What challenges are faced in performing query expansion in the clinical domain?**

  **Finding:** In this thesis, we explored using the two most commonly used ontologies to perform query expansion in the clinical and biomedical domains. In using MeSH, the system created relatively few expansions, but focused on the generation of synonyms. The second expansion system utilised the UMLS Metathesaurus, and although its expansive nature introduced significant noise, a more refined selection of expansion terms and relations in the UMLS Metathesaurus saw gains in performance to outperform MeSH based approaches.

- **What is the required performance of NLP, in order to aid IR approaches?**

  **Finding:** The Medical Entity Recognition system achieved exact f-scores up to 0.77. The Assertion Classification system posts results comparable to the state-of-the-art, while the concept normalisation system achieves an accuracy of 0.509. While the MER system was effective in annotating phrasal boundaries, it was still outperformed by a keyword baseline. Despite poor to moderate performance of the concept normalisation system on CLEF evaluations, it proved the most effective to the IR task. Assertion Classification, despite high performance on i2b2/VA evaluations, significantly hampers retrieval performance.
8.3 Future Research Considerations

In this thesis, we investigated the application of Clinical Natural Language Processing techniques to the Information Retrieval domain. In this section, we propose four avenues of research designed to extend work performed in this thesis.

8.3.1 State-of-the-art Concept Normalisation & Information Retrieval

Concept Normalisation extracts a span of tokens from free text, identified to be a medical concept that is then mapped to a unique identifier in an ontology, in this case the UMLS Metathesaurus. In the course of this thesis we developed and evaluated one such approach to concept normalisation. Concept Normalisation in Clinical Information Retrieval aims to address the issue of vocabulary mismatch which is commonplace in the clinical setting given its vast terminology and rich synonymy. The best performing normalisation approach in this thesis achieved an accuracy score of 0.513 on CLEF evaluation data. However, the official results of the CLEF/ShARe Evaluation Lab as in Table 3.18 show state-of-the-art approaches achieving accuracy measures of up to 0.939.

In spite of the moderate performance of the concept normalisation system presented in this thesis on CLEF evaluations, it has been shown in our experimentation to be effective enough to provide significant improvements to Clinical Information Retrieval approaches, outperforming all other NLP approaches presented in this thesis.

The improvement in Clinical Information Retrieval performance despite the low performance of the Concept Normalisation approach is actually converse to findings in the literature of the general textual news domain. For example, the equivalent task of Word Sense Disambiguation requires optimal performance (greater than 90% accuracy) to have any positive impact on retrieval performance (Sanderson, 1994). A possible contributory factor to this discrepancy in performance is the fact that as part of CLEF evaluations, only one CUI may be selected as a normalised concept. However, concepts in the UMLS are often very closely related and the can lead to issues whereby simply choosing one gold standard CUI is problematic. For example, given a concept text of liver failures we will extract two possible CUIs which map to Liver Failure, Acute and
Chronic Liver Failure. However, given the context of the text of liver failure we cannot decide whether the condition is acute or chronic. In an IR setting, it would be desirable to use both normalisations to maximise our recall of liver failures. However, current Concept Normalisation evaluation techniques take no account of such situations.

8.3.2 Supervised Evidence Combination Methods

In this thesis, we investigated several methods to improve the performance of Information Retrieval in the clinical domain. The top performing Information Retrieval system in this thesis combined keyword type queries with concept normalisation information, i.e. CUI’s from the UMLS. The motivation behind this was to investigate if the inadequacies of one approach may be counteracted by another. For example, while keyword retrieval approaches may facilitate good recall, their lack of term dependency information has been shown to be harmful to performance. In this case normalised concept information injected into a query could improve performance through ensuring that the correct concept is described in the text. However, the approach described assigns equal weighting to both keyword and conceptual queries, this allows the undesirable outcome whereby the poor performance of one approach may in fact be to the detriment of the other approach that may attain optimal performance for a given query.

To address this key problem area in this approach, a more selective method could be implemented. One such system is described in Limspotham et al. (2013) whereby the outputs of several Information Retrieval approaches are combined on a per query basis to achieve performance competitive with the state of the art. A supervised machine learning framework is applied to determine the weighting assigned to each approach prior to a combination of their outputs. This more novel approach to combining evidence allows the system to decide on a per query basis what retrieval approach contributes best to a given query. Two approaches may be leveraged in such approaches. The first may select a retrieval system e.g. keyword, phrasal, concept-based using criteria found within a query, e.g. numerous multi-word concepts. Another approach can use a voting based mechanism. Rather than discount a given approach entirely, a voting mechanism may be employed to lessen the impact of poorer performing approaches while increasing overall coverage. As is evidenced by our experimentation in Chapter 7, while the keyword and concept
based methods attain top performance, there are queries where one system is harmful to the other, for example in query 157 erroneous normalisation is to the detriment of keyword performance. For this reason, it would be an interesting direction for future work to investigate the use of a supervised model to combine the approaches described in this thesis.

8.3.3 Novel Query Expansion & Weighting Normalised Concepts

In this thesis, two ontologies, namely MeSH and the UMLS Metathesaurus were leveraged to perform query expansion. Query Expansion addresses the issue of vocabulary mismatch, whereby the terms in a document and query express the same key concepts, however the terms used to do so are different, i.e. they are synonyms of one another. The use of the MeSH system on average doubled the number of query terms while the vastly more expansive Metathesaurus generated nearly 20 times the number of query terms.

As shown in Section 7.4.2 the performance of the expansion technique was often determined by the concepts in the queries, rather than the technique itself. For example, in some queries MeSH outperformed a raw UMLS approach, in other queries the opposite is true. However, we note a generalisation in that MeSH had shortcomings in providing no expansions in some cases, while the UMLS approach tended to overgeneralise due to the high level of expansions. Thus we see opportunity in the use of the UMLS as opposed to MeSH, as we may harness its expansion power. Two runs which refined UMLS expansions to relations of synonymy, or narrower relations, e.g. tooth is narrower than dental, saw significant performance gains over the less expansive MeSH-based approach.

Given the promise of the UMLS as an expansion resource, we aim to leverage more novel methods of extracting expansion terms from the UMLS Metathesaurus. One such approach presented by Martinez et al. (2012) shows that the use of personalised Pagerank (Haveliwala, 2002) to select expansion terms from the UMLS Metathesaurus network graph can attain results competitive with the state-of-the-art retrieval systems. In this approach, query expansions are created through calculating personalised PageRank with respect to concepts specified in the query, with expansions generated through the UMLS. The system described in Martinez et al. (2012) identifies concepts in the queries using the MetaMap system, which has however been shown to have poor
concept identification abilities (Abacha and Zweigenbaum, 2011). In future work, we will investigate this personalised Pagerank method and leverage the concept normalisation system described in this thesis in order to further improve retrieval performance.

The more extensive use of the UMLS network structure also provides an avenue for future work in the weighting of normalised concepts. In this thesis, the top performing Information Retrieval system leverages concept normalisation. With this approach, all terms and concepts carry equal weights. This allows scenarios whereby a keyword which is too general, e.g. pain may have the same weight as a concept specifying tooth pain. Recent work has shown that the weighting of concepts can improve retrieval performance (Koopman et al., 2012b). In this approach, concepts may be weighted by their distance in the network to the original query term. The use of our concept normalisation system in combination with a semantic similarity weighting approach provides another possible area for future work.

8.3.4 Improving Clinical NLP performance

Though the performances of the NLP approaches discussed in this thesis are comparable to systems described in the literature, there is still room for improvement. Our work in this thesis focused on identifying features that contribute most to the task. Future work would investigate expanding the feature-set to achieve state-of-the-art performance.

Medical Entity Recognition

For the MER task, we restricted the size of our feature set to improve the tractability of training the system while investigating our key aim of discovering which features aid the task. However, state-of-the-art approaches leverage the complete feature sets of previous and successor tokens as well as the features of the current instance to create a more generalisable, though exponentially larger feature space (Tang et al., 2012). We have identified a key area for the improvement of MER, namely the recognition of disjoint entities as described as part of the CLEF/ShARe task. The system investigated in this thesis leveraged further semantic information in order to annotate tokens that occur within disjoint spans. However, further inspection of surface level features and pattern based features may improve disjoint recognition. For example, disjoint instances frequently occur in the pattern “[BODYPART] : [DESCRIPTION]” at the start of a line, where the combination of [BODYPART] and [DE-
SCRIBE] denote a disjoint entity separated by the span “:”, e.g. “Abdomen: Mild distension”. As the MER system discussed in this thesis primarily leveraged word-level features, a future focus on disjoint entity recognition will also include the use of further dependency tree features. An accurate dependency tree analysis can provide much more information about the structure of disjoint entities than surface features as they may be linked through a larger phrase structure.

**Assertion Classification**

While the multi-class classification performance of the Assertion Classification system in this thesis is comparable to state-of-the-art approaches, its application to the Information Retrieval task saw a significant introduction of errors. This shows that, despite high performance on shared tasks, the systems must be improved if they are to aid an Information Retrieval system. From the impressive ability of token-based and score-based features, we can identify that patterns play a large role in the classification of a medical conditions assertion status. Further to this, we have identified the key role that section heading information in reports plays in the assertion status of medical conditions, in particular *Family History* in the case of conditions experienced by someone else, and *Allergies* in the case of conditions that *may* affect the patient. Given the high frequency of these occurrences, a simple rule-based system may suffice in these exact cases. Therefore, a hybrid system which comprises of both manual rules as well as machine learning algorithms will be investigated in future work.
8.4 Summary

Government initiatives as well as cost and safety benefits as discussed in Chapter 1 have led to a growth in the adoption of Electronic Health Records over paper-based methods in healthcare institutions. This presents new avenues of research for the NLP and IR communities in developing methods to aid next-generation healthcare. While research in the clinical domain has been stifled in the past through data sharing constraints, new shared tasks such as the i2b2 and CLEF challenges as well as new repositories such as iDASH (Ohno-Machado et al., 2012) have seen research accelerate.

In this thesis, we explored the use of NLP methodologies to aid the Information Retrieval task in a clinical setting. Our findings in combining these approaches facilitate future work in both Clinical NLP and Clinical IR as interest grows and data becomes more widely available.
APPENDIX

A

TREC MEDICAL TRACK QUERY SET
<table>
<thead>
<tr>
<th>ID</th>
<th>Query Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>patients with hearing loss</td>
</tr>
<tr>
<td>102</td>
<td>patients with complicated GERD who receive endoscopy</td>
</tr>
<tr>
<td>103</td>
<td>hospitalized patients treated for methicillinresistant staphylococcus aureus MRSA endocarditis</td>
</tr>
<tr>
<td>104</td>
<td>patients diagnosed with localized prostate cancer and treated with robotic surgery</td>
</tr>
<tr>
<td>105</td>
<td>patients with dementia</td>
</tr>
<tr>
<td>106</td>
<td>patients who had positron emission tomography PET magnetic resonance imaging MRI or computed tomography CT for staging or monitoring of cancer</td>
</tr>
<tr>
<td>107</td>
<td>patients with ductal carcinoma in situ DCIS</td>
</tr>
<tr>
<td>108</td>
<td>patients treated for vascular claudication surgically</td>
</tr>
<tr>
<td>109</td>
<td>women with osteopenia</td>
</tr>
<tr>
<td>110</td>
<td>patients being discharged from the hospital on hemodialysis</td>
</tr>
<tr>
<td>111</td>
<td>patients with chronic back pain who receive an intraspinal pain medicine pump</td>
</tr>
<tr>
<td>112</td>
<td>female patients with breast cancer with mastectomies during admission</td>
</tr>
<tr>
<td>113</td>
<td>adult patients who received colonoscopies during admission which revealed adenocarcinoma</td>
</tr>
<tr>
<td>114</td>
<td>adult patients discharged home with palliative care home hospice</td>
</tr>
<tr>
<td>115</td>
<td>adult patients who are admitted with an asthma exacerbation</td>
</tr>
<tr>
<td>116</td>
<td>patients who received methotrexate for cancer treatment while in the hospital</td>
</tr>
<tr>
<td>117</td>
<td>patients with post-traumatic stress disorder</td>
</tr>
<tr>
<td>118</td>
<td>adults who received a coronary stent during an admission</td>
</tr>
<tr>
<td>119</td>
<td>adult patients who presented to the emergency room with anion gap acidosis secondary to insulin dependent diabetes</td>
</tr>
<tr>
<td>120</td>
<td>patients admitted for treatment of CHF exacerbation</td>
</tr>
<tr>
<td>121</td>
<td>patients with CAD who presented to the emergency department with acute coronary syndrome and were given plavix</td>
</tr>
<tr>
<td>122</td>
<td>patients who received total parenteral nutrition while in the hospital</td>
</tr>
<tr>
<td>123</td>
<td>diabetic patients who received diabetic education in the hospital</td>
</tr>
<tr>
<td>124</td>
<td>patients who present to the hospital with episodes of acute loss of vision secondary to glaucoma</td>
</tr>
<tr>
<td>125</td>
<td>patients coinfected with hepatitis C and HIV</td>
</tr>
<tr>
<td>126</td>
<td>patients admitted with a diagnosis of multiple sclerosis</td>
</tr>
<tr>
<td>127</td>
<td>patients admitted with morbid obesity and secondary diseases of diabetes and or hypertension</td>
</tr>
<tr>
<td>128</td>
<td>patients admitted for hip or knee surgery who were treated with anticoagulant medications postop</td>
</tr>
<tr>
<td>129</td>
<td>patients admitted with chest pain and assessed with CT angiography</td>
</tr>
<tr>
<td>130</td>
<td>children admitted with cerebral palsy who received physical therapy</td>
</tr>
<tr>
<td>131</td>
<td>patients who underwent minimally invasive abdominal surgery</td>
</tr>
<tr>
<td>132</td>
<td>patients admitted for surgery of the cervical spine for fusion or discectomy</td>
</tr>
<tr>
<td>133</td>
<td>patients admitted for care who take herbal products for osteoarthritis</td>
</tr>
<tr>
<td>134</td>
<td>patients admitted with chronic seizure disorder to control seizure activity</td>
</tr>
<tr>
<td>135</td>
<td>cancer patients with liver metastasis treated in the hospital who underwent a procedure</td>
</tr>
</tbody>
</table>

Table A.1: Topics from the 2011 TREC Medical Track
<table>
<thead>
<tr>
<th>ID</th>
<th>Query Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>136</td>
<td>Children with dental caries</td>
</tr>
<tr>
<td>137</td>
<td>Patients with inflammatory disorders receiving TNF-inhibitor treatments</td>
</tr>
<tr>
<td>138</td>
<td>Patients with acute tubular necrosis due to aminoglycoside antibiotics</td>
</tr>
<tr>
<td>139</td>
<td>Patients who presented to the emergency room with an actual or suspected miscarriage</td>
</tr>
<tr>
<td>140</td>
<td>Patients who developed disseminated intravascular coagulation in the hospital</td>
</tr>
<tr>
<td>141</td>
<td>Adult inpatients with Alzheimer disease admitted from nursing homes with pressure ulcers</td>
</tr>
<tr>
<td>142</td>
<td>Patients admitted with Hepatitis C and IV drug use</td>
</tr>
<tr>
<td>143</td>
<td>Patients who have had a carotid endarterectomy</td>
</tr>
<tr>
<td>144</td>
<td>Patients with diabetes mellitus who also have thrombocytopenia</td>
</tr>
<tr>
<td>145</td>
<td>Patients with lupus nephritis and thrombotic thrombocytopenic purpura</td>
</tr>
<tr>
<td>146</td>
<td>Patients treated for post-partum problems including depression hypercoagulability or cardiomyopathy</td>
</tr>
<tr>
<td>147</td>
<td>Patients with left lower quadrant abdominal pain</td>
</tr>
<tr>
<td>148</td>
<td>Patients acutely treated for migraine in the emergency department</td>
</tr>
<tr>
<td>149</td>
<td>Patients with delirium hypertension and tachycardia</td>
</tr>
<tr>
<td>150</td>
<td>Patients who have cerebral palsy and depression</td>
</tr>
<tr>
<td>151</td>
<td>Patients with liver disease taking SSRI antidepressants</td>
</tr>
<tr>
<td>152</td>
<td>Patients with Diabetes exhibiting good Hemoglobin A1c Control &lt; 8.0 %</td>
</tr>
<tr>
<td>153</td>
<td>Patients admitted to the hospital with end-stage chronic disease who are offered hospice care</td>
</tr>
<tr>
<td>154</td>
<td>Patients with Primary Open Angle Glaucoma (POAG)</td>
</tr>
<tr>
<td>155</td>
<td>Heart Failure (HF) Beta-Blocker Therapy for Left Ventricular Systolic Dysfunction (LVSD)</td>
</tr>
<tr>
<td>156</td>
<td>Patients with depression on anti-depressant medication</td>
</tr>
<tr>
<td>157</td>
<td>Patients admitted to hospital with symptomatic cervical spine lesions</td>
</tr>
<tr>
<td>158</td>
<td>Patients with esophageal cancer who develop pericardial effusion</td>
</tr>
<tr>
<td>159</td>
<td>Patients with cerebral edema secondary to infection</td>
</tr>
<tr>
<td>160</td>
<td>Patients with Low Back Pain who had Imaging Studies</td>
</tr>
<tr>
<td>161</td>
<td>Patients with adult respiratory distress syndrome</td>
</tr>
<tr>
<td>162</td>
<td>Patients with hypertension on anti-hypertensive medication</td>
</tr>
<tr>
<td>163</td>
<td>Patients treated for lower extremity chronic wound</td>
</tr>
<tr>
<td>164</td>
<td>Adults under age 60 undergoing alcohol withdrawal</td>
</tr>
<tr>
<td>165</td>
<td>Patients who have gluten intolerance or celiac disease</td>
</tr>
<tr>
<td>166</td>
<td>Patients who have hypoaldosteronism and hypokalemia</td>
</tr>
<tr>
<td>167</td>
<td>Patients with AIDS who develop pancytopenia</td>
</tr>
<tr>
<td>168</td>
<td>Patients with Coronary Artery Disease with Prior Myocardial Infarction on Beta-Blocker Therapy</td>
</tr>
<tr>
<td>169</td>
<td>Elderly patients with subdural hematoma</td>
</tr>
<tr>
<td>170</td>
<td>Adult patients who presented to the emergency room with suicide attempts by drug overdose</td>
</tr>
<tr>
<td>171</td>
<td>Patients with thyrotoxicosis treated with beta-blockers</td>
</tr>
<tr>
<td>172</td>
<td>Patients with peripheral neuropathy and edema</td>
</tr>
<tr>
<td>173</td>
<td>Patients over 65 who had Pneumonia Vaccination Status presently or previously</td>
</tr>
<tr>
<td>174</td>
<td>Elderly patients with ventilator-associated pneumonia</td>
</tr>
<tr>
<td>175</td>
<td>Elderly patients with endocarditis</td>
</tr>
<tr>
<td>176</td>
<td>Patients with Heart Failure (HF) on Angiotensin-Converting Enzyme (ACE) Inhibitor or Angiotensin Rec</td>
</tr>
<tr>
<td>177</td>
<td>Patients treated for depression after myocardial infarction</td>
</tr>
<tr>
<td>178</td>
<td>Patients with metastatic breast cancer</td>
</tr>
<tr>
<td>179</td>
<td>Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression</td>
</tr>
<tr>
<td>180</td>
<td>Patients with cancer who developed hypercalcemia</td>
</tr>
<tr>
<td>181</td>
<td>Patients being evaluated for secondary hypertension</td>
</tr>
<tr>
<td>182</td>
<td>Patients with Ischemic Vascular Disease</td>
</tr>
<tr>
<td>183</td>
<td>Patients presenting to the emergency room with acute vision loss</td>
</tr>
<tr>
<td>184</td>
<td>Patients with Colon Cancer who had Chemotherapy</td>
</tr>
<tr>
<td>185</td>
<td>Patients who develop thrombocytopenia in pregnancy</td>
</tr>
</tbody>
</table>

Table A.2: Topics from the 2012 TREC Medical Track


180


Jiang, M., Chen, Y., and Liu, M. (2010). Hybrid approaches to concept extraction and assertion classification - vanderbilt’s systems for the 2010 i2b2 nlp challenge. In *Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*.


