Semantic tagging of medical narratives using SNOMED CT

Saman Hina

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy in Computing

The University of Leeds
School of Computing

August, 2013.

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Acknowledgements

I would like to thank my supervisor Dr. Eric Atwell for his valuable understanding, patience and most importantly his guidance throughout my studies. I would also like to thank my co-supervisor Owen Johnson for providing practical opportunities and guidance in the field of healthcare research.

Very special thanks to my mentor, Dr. Katja Markert for her valuable suggestions and critical feedback on my work.

Thanks to all my friends and colleagues, especially Dr. Krishna Sandeep Dubba Reddy, Sammy Danso, Niraj Aswani, Samantha Crossfield, Dr. Claire Brierley and Dr. Abdul Baqi Sharaf for their support, useful feedback and discussions. I would also like to thank Dr. Richard Jones from Leeds Institute of Health Sciences and Dr. Marc Jamoulle for reviewing the gold standard. Thanks to my friends Momina Khan and Amara Raja for annotating the gold standard.

I would like to thank the 2010 i2b2/VA challenge organizers for providing the corpus from different healthcare partners. The 2010 i2b2/VA challenge and the workshop were funded in part by the grant number U54-LM008748 on Informatics for Integrating Biology to the Bedside from National Library of Medicine. This challenge and workshop were also supported by resources and facilities of the VA Salt Lake City Health Care System with funding support from the Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374 and the VA Informatics and Computing Infrastructure (VINCI), VA HSR HIR 08-204. MedQuist, the largest transcription technology and services vendor, generously co-sponsored the 2010 i2b2/VA challenge meeting at AMIA. Thanks to TTP Ltd for providing test data and my employer ‘NED University of Engineering & Technology’ for funding my research.

Most importantly, I would like to thank my parents Mr Muhammad Taujeeh-ul-Haque (Late) and Mrs Nasima Taujeeh who really encouraged my academic career and my sister GulRukh Hina for taking care of my mother in my absence.

I would also like to thank my parents in law for their support and patience. Finally, thanks to my loving husband Mr Muhammad Ali; the generation of this thesis would not have been possible without the friendship, support and encouragement he has provided me during the course of my studies.

Abstract

In the medical domain, semantic analysis is critical for several research questions which are not only limited to healthcare researchers but are of interest to NLP researchers. Yet, most of the data exists in the form of medical narratives. Semantic analysis of medical narratives is required to be carried out for the identification of semantic information and its classification with semantic categories. This semantic analysis is useful for domain users as well as non-domain users for further investigations.

The main objective of this research is to develop a generic semantic tagger for medical narratives using a tag set derived from SNOMED CT® which is an international healthcare terminology. Towards this objective, the key hypothesis is that it is possible to identify semantic information (paraphrases of concepts, abbreviations of concepts and complex multiword concepts) in medical narratives and classify with globally known semantic categories by analysis of an authentic corpus of medical narratives and the language of SNOMED CT®.

This research began with an investigation of using SNOMED CT® for identification of concepts in medical narratives which resulted in the derivation of a tag set. Later in this research, this tag set was used to develop three gold standard datasets. One of these datasets required anonymization because it contained four protected health information (PHI) categories. Therefore, a separate module was developed for the anonymization of these PHI categories. After the anonymization, a generic annotation scheme was developed and evaluated for the annotation of three gold standard datasets. One of the gold standard datasets was used to develop generic rule-patterns for the semantic tagger while the other two datasets were used for the evaluation of semantic tagger. Besides evaluation using the gold standard datasets, the semantic tagger was compared with three systems based on different methods, and shown to outperform them.

Declarations

Much of the work presented in this thesis has been published in the following research papers. Detail of each publication with respect to chapters is given below;

Chapter 3

HINA, S., ATWELL, E., JOHNSON, O. & BRIERLEY, C. (2013) "Identification, Classification and Anonymisation of 'Protected Health Information' in real-time medical data for research purposes", presented in: *The 23rd Meeting of Computational Linguistics in the Netherlands (CLIN 2013).* Netherlands.

Contributions: The work in this paper was contributed and written by Hina, S.

Atwell, E., Johnson, O. and Brierley, C. provided supervision, feedback and general guidance.

SMITH, R., XU, J., HINA, S. & JOHNSON, O. (2013) "GATEway to the Cloud Case Study: A privacy-aware environment for Electronic Health Records research", published in: *IEEE MobileCloud2013 Industry Track in conjunction with SOSE 2013*, San Francisco Bay, USA.

Contributions: Smith, R. worked on cloud based module and contributed to paper write-up. Hina, S. developed NLP algorithms for anonymisation and helped in writing the relevant module.

Johnson, O. and Xu, J. provided supervision, feedback and general guidance.

Chapter 4

HINA, S., ATWELL, E. & JOHNSON, O. (2010) "Secure Information Extraction from Clinical Documents Using SNOMED CT Gazetteer and Natural Language Processing", In Proceedings of: *The 5th International Conference for Internet Technology and Secured Transactions (ICITST-2010).*

Contributions: The work in this paper was contributed and written by Hina, S.

Atwell, E. and Johnson, O. provided supervision, feedback and general guidance.

HINA, S., ATWELL, E. & JOHNSON, O. (2010) "Semantic Tagging of Medical Narratives with Top Level Concepts from SNOMED CT Healthcare Data Standard", published in: *International Journal of Intelligent Computing Research (IJICR),* 1**,** 118-123.

Contributions: The work in this paper was contributed and written by Hina, S.

Atwell, E. and Johnson, O. provided supervision, feedback and general guidance.

Chapter 5

HINA, S., ATWELL, E., JOHNSON, O. & WEST, R. (2010) "Extracting the concepts in Clinical Documents using SNOMED-CT and GATE", In: *Fourth i2b2/VA Shared-Task and Workshop, Challenges in Natural Language Processing for Clinical Data.*

Contributions: The work in this paper was contributed and written by Hina, S.

Atwell, E. and Johnson, O. provided supervision, feedback and also contributed in reviewing the paper.

HINA, S., ATWELL, E. & JOHNSON, O. (2011) "Enriching a healthcare corpus with SNOMED CT standard medical semantic tags", In Proceedings of: *Corpus Linguistics Conference 2011, Discourse and Corpus Linguistics.* Birmingham.

Contributions: The work in this paper was contributed and written by Hina, S.

Atwell, E. and Johnson, O. provided supervision, feedback and also contributed in reviewing the paper.

HINA, S., ATWELL, E. & JOHNSON, O. (2012) "Development and evaluation of annotation guidelines for non-domain- experts for a gold standard corpus of medical narratives", Submitted in: *International Journal of Corpus Linguistics.*

Contributions: The work in this paper was contributed and written by Hina, S.

Atwell, E. and Johnson, O. provided supervision, feedback and general guidance.

Chapter 6 and Chapter 7

HINA, S., ATWELL, E. & JOHNSON, O. (2012) "Automated analysis of domain specific corpus in healthcare domain: For non-domain users". Paper presented in: *The Sixth Inter-Varietal Applied Corpus Studies (IVACS) International Conference*. United Kingdom.

Contributions: The work in this paper was contributed and written by Hina, S.

Atwell, E. and Johnson, O. provided supervision and feedback. Atwell, E. also contributed in reviewing the paper.

HINA, S., ATWELL, E. & JOHNSON, O. (2013) "SnoMedTagger: A semantic tagger for medical narratives", In: *14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013)*, University of the Aegean, Samos, Greece.

Contributions: The work in this paper was contributed and written by Hina, S.

Atwell, E. and Johnson, O. provided supervision, feedback and general guidance.

HINA, S., ATWELL, E. & JOHNSON, O. (2013) "SnoMedTagger: A semantic tagger for medical narratives", to be published in: *International Journal of Computational Linguistics and Applications (IJCLA).*

Contributions: The work in this paper was contributed and written by Hina, S.

Atwell, E. and Johnson, O. provided supervision, feedback and general guidance.

Table of Contents

List of Figures

List of Tables

Abbreviations

ANNIE – A Nearly New Information Extraction

CREOLE – Collection of Reusable Objects plugins for language engineering

CPSL – Common Pattern Specification Language

EHR – Electronic Health Record

GATE – General Architecture for Text Engineering

HIPAA – Health Information Portability and Accountability Act (1996)

HUGO – Human Genome Organisation

IAA – Inter Annotator Agreement

ICD – International Classification of Diseases

IE – Information Extraction

JAPE – Java Annotation Pattern Engine

ML – Machine Learning

NER – Named Entity Recognition

NLP – Natural Language Processing

PHI – Protected Health Information

RRF – Rich Release Format

SNOMED CT® – Systemised Nomenclature of Medicine – Clinical Terms

SVM – Support Vector Machines

UMLS – Unified Medical Language

VRE – Virtual Research Environment

XML – Extensible Markup Language

Important definitions

Anonymization: The process in which data fields that may be used to identify the individuals to whom the data records relate are removed from a data set.

Corpus: Collection of documents/texts.

De-identification: The term de-identification refers to removing identifiers from data without losing the linkage of hidden identifiers.

Gold standard corpus: A corpus that contains the identified/annotated information. In Natural Language Processing applications, gold standard corpus is required for evaluation the performance of automatic system against gold standard annotations.

Metadata: Data about data is called metadata.

Named Entity Recognition: Anything that can be referred to by a proper name is a 'Named Entity'. The process that identifies proper names in the text and classifies them with respective named entities is known as 'Named Entity Recognition'.

Natural Language Processing: Natural language processing (NLP) is a field of 'Artificial Intelligence'. NLP is the ability of a computer program/application to understand natural language.

Semantics: The study or science of meaning/ interpretation of language.

Semantic Analysis: A process that determines which words or phrases in the text are relevant to the domain and then assigns their semantic relations.

Semantic Tagging: The identification of semantic information in the text and its classification with respective semantic categories is known as semantic tagging.

SnoMedTagger: SnoMedTagger - SNOMED CT Medical Tagger is a generic semantic tagger that was developed specifically for tagging semantic information medical narratives using semantic categories derived from an international healthcare terminology, SNOMED CT®.

Tagging: The identification of required information and its association with respective tag/category/type is called tagging.

Chapter 1. Introduction

In Natural Language Processing, the term ‘Semantics’ represents the study of the meaning of a language. More specifically, semantics has potential use in the investigation of a number of research questions that are related to language (Jurafsky and Martin 2009). The identification of semantic information in text and its classification with respective semantic categories is known as semantic tagging. Semantic tagging enriches information to improve the analysis of text in a given domain. Semantic tagging can be carried out on spoken and written language including the technical language which is used in a specialised domain such as the medical domain, law, chemistry and so on.

This research deals with one specialised domain - the medical domain. In the medical domain, much of the data exists in the form of medical narratives written by clinicians in the form of unstructured free text. This unstructured data resides in Electronic Health Record (EHR) systems. This data is a result of data entry of manual records in EHR systems, transcriptions of dictations by radiologists or using speech recognition software for recording consultations. This unstructured form (medical narratives) may suit the individual human reader who can interpret the subtlety of the language and use it to inform their clinical decision making, but it is difficult for searching, analysing and understanding the meaning of concepts or terms that are present in the medical narratives. Thus, NLP is needed to identify and classify important semantic information (concepts) within the medical narratives for more structured analysis (Meystre 2008). The semantic tagging of the data is a necessary step in the process of using medical narratives to inform many research tasks such as ‘finding cause of death’, ‘extracting diagnoses and so on. The following section explains the identification and classification of semantic information (semantic tagging) in the medical domain.

1.1 Semantic tagging in the medical domain

In the medical domain, clinicians (domain experts in the context of this study) record their consultations and other clinical documents in Electronic Health Record (EHR) systems. For this

purpose, they use a combination of structured information, coded data, and medical narratives using natural language, also referred to as unstructured text. Clinical documents such as discharge summaries, progress notes, and medical reports contain important information which needs to be shared for research purposes. Where natural language is used in clinical documents, the semantic information varies from one clinician to another. This is because of differences in the expressiveness of language, the use of synonyms, paraphrases, abbreviations, etc. These variations in natural language free text follow informal writing structure and can therefore obscure important information within text. The result is that the researchers may find the narrative confusing, ambiguous or imprecise and this can potentially lead to misunderstanding. As a result, some crucial information might not be extracted from the text. In such situations, the identification and classification of semantic information (semantic tagging) can facilitate a more consistent interpretation of the natural language written by clinicians. The approach may also help researchers in dealing with research questions that cannot be answered by analysis of the structured and coded elements of EHRs.

In the medical domain, researchers who use medical narratives in their research usually hire domain experts to identify and classify the semantic information within the natural language, a process which is time consuming and expensive. This means that non-domain users (such as language researchers) are dependent on domain experts to identify and classify semantic information. The process of 'annotation', i.e., the identification and classification of semantic information with respective semantic categories can be automated using a computerised system, typically referred to as a 'semantic tagger'. In Computational Linguistics, a 'semantic tagger' is the term used in 'Information Extraction' (IE) applications. Another IE application called 'Named Entity Recognition' is closely related to semantic tagging. The difference between these two applications is that named entity recognition applications only identify and classify *proper names* in the text while semantic tagging identifies and classifies semantic metadata (data/information about data) in the text.

This research study dealt with the development of a generic semantic tagger which can be employed for extraction of semantic information in medical narratives. The developed semantic tagger was named SnoMedTagger - SNOMED CT Medical Tagger (available at http://www.comp.leeds.ac.uk/scsh/SnoMedTagger.html) and it uses the semantic categories derived from an international healthcare clinical terminology SNOMED CT® or Systemised Nomenclature of Medicine - Clinical Terms. SNOMED CT® is globally the most comprehensive clinical terminology and it is specified in several US standards (Stearns et al. 2001). SNOMED CT healthcare terminology and its components are described in detail in Chapter 4.

In this study, the semantic metadata of interest in medical narratives are the 'concepts' or clinical terms that can be classified into appropriate semantic categories. For instance, 'CT Scan' and 'lungs' belong to the semantic categories 'Procedure' and 'Body Structure', respectively. The metadata present in medical narratives can be in the form of individual concepts, paraphrases of concepts, abbreviations of concepts and complex multiword concepts. Chapter 2 will explore the discussion above in more detail using the more technical language of Natural Language Processing.

1.2 Motivation and goals for this research

In the medical domain a significantly large proportion of the data in medical records is in the form of medical narratives. This is because it is often preferred by clinicians as a way of recording patient health information due to its richness and convenience. However, the analysis of the semantic information within these medical narratives is more complex as a result (illustrated in Section 1.1).

When non-domain researchers such as NLP researchers work on a particular research question that involves the use of medical narratives, they typically hire domain experts for the annotation of the required information to create a gold standard data (annotated information). The primary limitation of this approach is that it may restrict the annotated data to specific research task and/or question and limit more general use. This is because different researchers working on

medical narratives use different names for synonymous semantic categories. For instance, the semantic category 'Test' can also be referred to as a 'Procedure' or the semantic category 'Treatment' can also be named as 'Medications'. In addition, the names of various semantic categories may or may not necessarily be the same as those used in various healthcare clinical terminologies. This research recognised the need for the development of a generic semantic tagger for medical narratives based on standard semantic categories derived from an international healthcare clinical terminology. In this case, such a system (semantic tagger) could reduce or even eliminate the need to employ domain experts when non-domain users (such as NLP researchers) analyse clinical documents in their research.

In addition to this, the use of semantic categories which are derived from SNOMED CT® could facilitate consistent information exchange between researchers whether they are domain users or not. The underlying hypothesis is that it is possible to identify and classify semantic information in medical narratives by developing generic rule-patterns derived from the following resources:

- An authentic corpus of medical narratives written by clinicians.
- The language of healthcare terminology SNOMED CT®.

The main contribution of this research has been to test this hypothesis by building a product to implement and refine a semantic tagger for medical narratives based on the classification structures in SNOMED CT®. The resulting product has been named ‘SnoMedTagger’ and is described in Chapter 6. Other challenges were tackled as secondary contributions and these are explained in the next section.

1.3 Contributions of this research

Primary Contribution: SnoMedTagger – a semantic tagger for medical narratives using SNOMED CT®

As described in Section 1.1, the identification and classification of semantic information is a pre-processing step for a range of research questions that involve the use of medical narratives.

For this purpose, domain experts can be hired but this approach suffers from the following major drawbacks.

- The process is expensive and time consuming.
- The identified semantic categories are inconsistent and are limited to the specific research question. Therefore, the identified semantic information cannot be used in dealing with other research questions.

To overcome these limitations, a generic semantic tagger named SnoMedTagger was developed in this work. The SnoMedTagger uses a medical semantic tag set of 16 semantic categories derived from SNOMED CT® health care terminology (Hina, Atwell and Johnson 2013b). The extraction of semantic categories from SNOMED CT is described in Chapter 4, while the development of SnoMedTagger is explained in Chapter 6. Due to the fact that SNOMED CT® is a comprehensive healthcare terminology for the exchange of information (*SNOMED CT User Guide, January 2011 International Release*) and it is also approved by the National Health Service in England (*NHS-Connecting for Health*), the semantic categories used in the SnoMedTagger are expected to be useful to domain users as well as non-domain users (Hina, Atwell and Johnson 2012).

The output of SnoMedTagger on a sample text is shown in Figure 1-1. The SnoMedTagger was able to identify and classify semantic information with respective categories. However, abbreviation of concept 'PTX' was an exception. This was due to the fact that this abbreviation was not found in the original SNOMED CT vocabulary. Moreover, the annotators also did not assign any semantic category to this concept abbreviation in the gold standard dataset.

While employing the SnoMedTagger for extraction of semantic information, the user can select only those semantic categories that are appropriate for their research task/question, as shown in Figure 1-1. Different colours can be chosen to differentiate between semantic categories and to avoid any confusion in colour coded output, the output can also be exported to XML format.

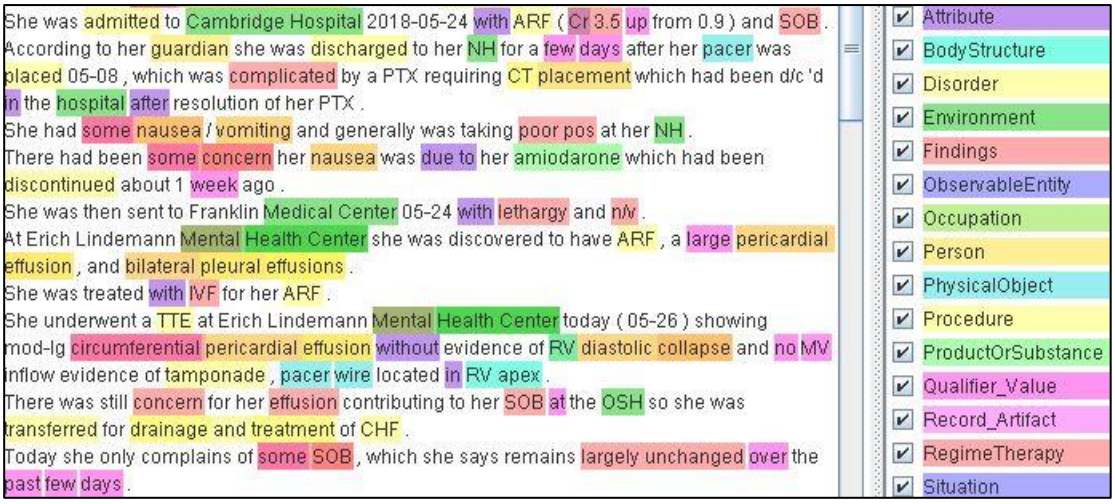


Figure 1-1: An example of the output of SnoMedTagger.

Secondary Contributions:

1) **Evaluation and validation of SnoMedTagger** – The existing well-known semantic tagging systems such as MetaMap (Aronson and Lang 2010) and onotology-based BioPortal web annotator (Noy et al. 2009) have not been evaluated on a gold standard dataset. The Metamap is considered as state-of the-art system in medical domain (Abacha and Zweigenbaum 2011). However, the SnoMedTagger was evaluated against two different gold standard datasets; Test dataset 1 and Test dataset 2. This was done to test the general applicability of rule-based SnoMedTagger on different medical narratives. Results of Test dataset 1 have been published in (Hina, Atwell and Johnson 2013b). This was followed by the validation of SnoMedTagger by two domain experts. It was reported that the SnoMedTagger, which is a rule-based system, outperformed the systems that are based on different methods/approaches; 1) SNOMED CT dictionary application: baseline system, 2) An Ontology-based 'BioPortal' web annotator and 3) SVM-based machine learning system (SVM - Support Vector Machine is a supervised machine learning classifier).
2) **Anonymization module for Test dataset 1 (Explained in Chapter 3)** – In the medical domain, data that contain Protected Health Information (PHI) about individuals require anonymization. This is due to ethical issues that are associated with the use of such data.

The three datasets which were used in the development and evaluation of SnoMedTagger are; the Development dataset, the Test dataset 1 and the Test dataset 2.

The Development dataset and the Test dataset 2 included de-identified/anonymized discharge summaries and progress notes which were accessed after ethical approval from the data providers. However, the Test dataset 1 which was obtained from an Electronic Health Record system known as 'SystmOne', mainly contained fictional information about individuals with some bits of real-data in it and therefore needed to be anonymized. In addition, the data contained a mixture of natural language and clinical codes and its characteristics were similar to any real data. Thus, an anonymization module was developed to anonymize the Test dataset 1 which can be used for the anonymization of real-data in SystmOne (Hina et al. 2013). This anonymization module also formed part of the 'e-Health Gateway to the Clouds' project. The objective of this project was to make authentic healthcare data available for research within a secure cloud-based VRE - Virtual Research Environment after anonymization (Smith et al. 2013). This module can be downloaded from http://www.comp.leeds.ac.uk/scsh/. For the fulfilment of ethical requirements, the anonymization is an essential pre-processing module for the SnoMedTagger in case of data containing PHI.

3) **General annotation guidelines for medical narratives –** For the development and evaluation of the rule-based SnoMedTagger, annotation of gold standard datasets (Development dataset, Test dataset 1, Test dataset 2) was required. For this purpose, simple and generic annotation scheme guidelines were developed for the annotation of semantic information (i.e. paraphrases of the concepts, abbreviations of the concepts, complex multiword concepts). These annotation guidelines were developed by considering the language issues that cannot be tackled using dictionaries or thesauri (Hina, Atwell and Johnson 2011).

1.4 Thesis structure

The structure of this thesis is as follows;

Chapter 2 includes a review of the work done by other researchers on semantic tagging in the medical/biomedical domain using different methods and resources.

Chapter 3 presents the development and evaluation of the anonymization module (secondary contribution 2). This module is not directly linked to the main contribution of this research, therefore instead of including its related work in Chapter 2 (Background chapter on semantic tagging); a complete section is included in this chapter.

Chapter 4 explains the use of SNOMED CT healthcare clinical terminology. The development of the baseline system (SNOMED CT dictionary application) using dictionaries of semantic categories derived from SNOMED CT is also described in this chapter.

Chapter 5 deals with the datasets that were used in this research and the annotation guidelines developed for the annotation of the gold standard datasets (secondary contribution 3). The annotation experiments that were conducted using the developed annotation guidelines and its evaluation are also explained in this chapter.

Chapter 6 presents the development of the rule-based semantic tagger (SnoMedTagger) which is the main contribution of this research.

Chapter 7 is regarding the evaluation and validation of the performance of SnoMedTagger against two different unseen gold standard test datasets (secondary contribution 1).

Chapter 8 contains a summary of the results achieved in this research. It also includes the limitations and the suggested future work.

Chapter 2. Background

This chapter contains a description of semantic tagging in general and the different approaches adopted by researchers for semantic tagging/annotation of texts in medical/biomedical domain. This review provided a basis for the development of SnoMedTagger, which is an NLP application for tagging semantic information in medical narratives.

2.1 A brief overview of semantic tagging

A corpus can be simply defined as a collection of texts. Tagging or annotation of a corpus (a collection of texts) is the process of adding tags to information in the corpus. In other words, tagging is an inline addition of respective category to the words in the corpus. Different types of tagging that are done in language research include *part-of-speech tagging* (Leech, Garside and Atwell 1983; Brill 1992; Atwell 2008; Sawalha and Atwell 2013), *syntactic tagging* (Zhou and Huang 1994; Dukes, Atwell and Habash 2013; Atwell et al. 2000; Atwell 1983) and *semantic tagging* (Demetriou and Atwell 2001; Huang et al. 2005; Brierley et al. 2013; Danso et al. 2013). The term 'semantic tagging' refers to an information extraction process that enriches information for better analysis of text in a given domain.

For instance, (Rau 1991) implemented an heuristic algorithm for extraction of 'company names' from financial news stories. This algorithm was not only able to extract company names but also their semantic variation. (Demetriou and Atwell 2001) used Longman English Dictionary Online (LDOCE) for semantic tagging of general English text. A different approach was adopted by (Boufaden 2003) based on domain specific ontology. They developed an ontology-based domain specific semantic tagger which focused on tagging semantic information in transcribed telephone conversations using concepts from a Search and Rescue ontology.

Another semantic tagger was included in the GATE (General Architecture of Text Engineering) software tool. The semantic tagger in the GATE was developed using JAPE - Java Annotation Pattern Engine rules (Cunningham, Mayard and Tablan 2000). JAPE rules are further described

in Section 6.2. In this semantic tagger, JAPE rules were developed for identification and classification of important semantic information in the text such as ‘date’, ‘organisation’, ‘location’, etc., (Cunningham et al. 2002). (Nadeau, Turney and Matwin 2006) developed an unsupervised system for extracting the classical categories (such as date, location) as well as domain specific semantic category, ‘car brands’.

(Popov et al. 2003) proposed an innovative model for automatic semantic annotation based on ontology and a massive knowledge base. The ontology contained general entities on upper-level and domain specific entities on lower-level in hierarchy. Therefore, this type of semantic annotation was able to provide information of general named entities such as Person, Location, Organisation, etc., as well as domain specific entities such as private organisations, public organisations, etc. This method can be used to improve semantic enrichment in documents. However, may increase the processing time depending on the annotation level.

Similarly, other researchers also reported their work on semantics using different approaches such as ontologies, rule-based and machine learning for identification and classification of semantic information using different type of texts (Yu-Chieh Wu et al. 2006; Kirchner and Sinot 2007; Christensen et al. 2009).

In the medical domain, semantic tagging of data was carried out in several investigations. Semantic tagging can be carried out for the development of an evaluation corpus. For instance, (Ogren, Savova and Chute 2008) annotated only the semantic category ‘Disorder’ using the SNOMED CT ontology. To develop an automatic CLEF (Clinical E-Science Framework) entity recognition system, semantic annotation was done by (Roberts A 2007) on CLEF corpus. The corpus contained histopathology reports, imaging reports and clinical narratives. In this project, researchers developed specific annotation schema for semantic entities (condition, intervention, investigation, result, drug or device, locus.) and their relationships (has_target, has_location, has_indication, has_location, co-refers, modifies [literality], modifies [sub-location], and modifies [negation]). This corpus is not publically available for research.

Since the present study dealt particularly with semantic tagging of medical/biomedical text, a more detailed account of relevant methods/approaches is presented in the next sections.

2.2 Semantic tagging using ontologies or dictionaries

Thesauri or ontologies are often used in the biomedical/medical domain. The use of ontologies provides synonyms (concepts/terms), hypernyms (in the hierarchy) and indexing (codes). Ontology-based and dictionary-based methods are usually simpler in implementation. However, the systems based on these methods cannot be successfully applied on medical narratives. This is due to limited expressiveness of language that is found in ontologies. Ontologies or terminologies such as Unified Medical Language - UMLS® (Lindberg, Humphreys and McCray 1993) are useful in extracting lexical knowledge but they do not include variations of phrases that occurs in medical narratives.

(Krauthammer et al. 2000) implemented a method based on BLAST algorithm that searches gene names in a database. It provides approximate matches and identifies small variation in gene names. They developed an automatic system for the identification of gene and protein names in journal articles. It is instructive to mention here that maintaining and updating such dictionaries are not easy tasks. For instance, (Hirschman et al. 2003) reported addition and withdrawal of 166 names in the Mouse Genome database[1] within a week.

Another approach based on dictionaries was presented by (Hanisch et al. 2003). They used a dictionary of gene and protein names for semantic classification in scientific literature. Their focus was on the automatic generation of dictionaries by extraction of symbols, aliases and gene names from HUGO Nomenclature (Wain et al. 2002) and their corresponding names from OMIM database[2]. In similar work, the synonyms of protein names were extracted from SWISSPROT and TREMBL databases. The extracted dictionary was then cured and pruned by resolving ambiguity issues and by generating more synonyms from dictionary terms. They

[1] http://www.informatics.jax.org/mgihome/nomen/short_genes.html
[2] http://www.ncbi.nlm.nih.gov/omim

calculated ‘specificity’ and ‘sensitivity’ for evaluation. Specificity measures the true negative rates (correctly rejected) while sensitivity measures the true positive rates (correctly identified/recall). Their semi-automatic approach of creating generic dictionary for the identification of gene and protein names with their synonyms achieved 95% specificity and 90% sensitivity on the corpus of MEDLINE abstracts. MEDLINE abstracts are structured articles; therefore the work done by these researchers did not guarantee its applicability on unstructured medical narratives.

(Long 2005) used SNOMED CT healthcare clinical terminology for coding semantic information (‘diagnosis’, ‘procedure’) extracted from a small corpus (23 documents) of discharge summaries. They used simple natural language processing to locate section headers of documents and then identify concept phrases that maps with SNOMED CT concepts in the UMLS (*Unified Medical Language System® (UMLS®)*). The limitation of this approach is that it has been developed for a small set of discharge summaries that contained clues of section headers such as punctuation marks and cannot be applicable on any other format. In addition to this limitation, these researchers did not assure the applicability of this method on other data because it was not tested on any data. Similarly, (Ogren, Savova and Chute 2008) used SNOMED CT healthcare clinical terminology for the development of a gold standard dataset that contained 1556 concept annotations. This gold standard dataset was used to evaluate their biomedical named entity recognition system. This corpus was taken from Mayo clinic repository which consists of clinical documents transcribed by clinicians. 82,813 'Disorder' concepts were extracted from the SNOMED CT healthcare terminology to annotate the semantic category of 'Disorder'. Four annotators annotated corpus of 47,975 words with the 'Disorder' semantic category, concept code and context. Then, the annotators used RRF

Browser[3] to search concepts by keyword or hierarchical navigation for annotation. The following two strategies were adopted to facilitate the annotators.

1. Two annotators were provided with a corpus that was already annotated using MetaMap system. However, the annotators were allowed to add or remove annotations following the annotation guidelines. This approach facilitated quick review and correction of annotations.

2. Using the same annotation guidelines, the other two annotators manually annotated the corpus without any pre-processing. This was done to verify the annotation guidelines.

In both strategies, annotators annotated the corpus independently. The consensus set was created for both cases and the final set was mutually completed by four annotators reviewing consensus sets achieved from both strategy 1 and 2. The overall agreement between the two consensus sets was 74.6%.

A semi-automatic tool called 'Semantator' was developed for annotating medical narratives (Song, Chute and Tao 2011). Semantator is a protégé plugin which allows manual annotation and semi-automatic annotation. In manual annotation, a user can annotate a piece of text using a class from the ontology loaded in protégé. Semi-automatic approach uses semantic web ontologies from BioPortal (Noy et al. 2009) and clinical Text Analysis and Knowledge Extraction System – cTAKES (Savova et al. 2010). The major drawback of this system is that it was not evaluated using any gold standard corpus of medical narratives and the gold standard was annotated with only one semantic category. Furthermore, the other limitations reported in this research are based on limited user experiences (Song, Chute and Tao 2012).

An automatic system for the analysis of semantic information in biomedical reports was developed by (Hahn, Romacker and Schulz 2002). This system used a domain specific lexicon and performed syntactic analysis on the basis of lexical definitions and dependency grammars.

[3] https://www.nlm.nih.gov/research/umls/implementation_resources/metamorphosys/RRF_Browser.html

With the help of a parser, grammatical constructions of lexical items found in the text were analysed. The parsed information in the text that helped in the derivation of concepts was then enriched with semantic annotation. The semantic annotation of text was achieved by automatic transformation of text into description logics format which was then mapped with a medical knowledgebase.

Another approach was adopted by (Baud, Rassinoux and Scherrer 1992) for the domain of 'digestive surgery'. They studied the representation of clinical narratives using conceptual graphs that were generated from single words in semantic lexicon and then used to form full sentences (Baud et al. 1995). This NLP system which is based on proximity parsing, allows browsing and encoding of concepts. In addition, the system is capable of handling multilingual data. However, the limitation is of being developed for specific domain (digestive surgery).

(Albright et al. 2013) reported manual annotation of syntactic and semantic information in clinical narratives. Semantic annotation was done using semantic groups instead of semantic categories, to avoid any confusion between the synonymous semantic categories in clinical narratives. This research involved the use of UMLS schema for semantic annotation of the following semantic groups; 'Procedure', 'Disorder', 'Concept and Ideas', 'Anatomy', 'Chemical and Drugs' and only one UMLS semantic category 'Sign or Symptom'. The corpus was pre-annotated with UMLS entities using clinical Text Analysis and Knowledge Extraction System – cTAKES (Savova et al. 2010). 74% of the corpus was double annotated by two annotators and the rest of 26% was single annotated. The double annotated data was then compiled to create the gold standard dataset. The inter-annotator agreement (IAA) was calculated using F-measure by considering the annotations of the first annotator as gold standard. For exact matches with UMLS concepts, 69.7% of IAA was reported and 75% IAA was achieved for partial matches.

Systems such as MetaMap (Aronson 2001) and BioPortal web annotator (Noy et al. 2009) also use ontologies for identification and classification of semantic categories. Since these systems use a number of ontologies, a major drawback is the potential of ambiguity of semantic

categories. To extract the semantic information in medical/biomedical text, MetaMap uses ontologies with extension of special modules based on regular expressions rules. Metamap was developed using MEDLINE abstracts containing structured journal articles. Therefore, Metamap is inappropriate for use on unstructured medical narratives (Patterson, lgo and Hurdle 2010). On the other hand, BioPortal web annotator contains more than 200 ontologies which can be used for the identification and classification of required semantic categories in the text (Noy et al. 2009). The BioPortal system is not suitable for semantic tagging of medical narratives because of limited language of ontologies. This point is established in Chapter 5.

In summary, in the context of the study reported in this thesis, the limitations of ontology-based or dictionary-based approaches include the following.

1. Limited language of ontologies.
2. Inconsistency of semantic information (semantic categories) used for different datasets.

It is proposed that the above mentioned limitations can be covered by applying rules or patterns on the output of dictionaries or ontologies. Rule-based or pattern-based methods (explained in the next section) provide better options in case of a small amount of annotated data because other methods, such as machine learning, require large annotated data.

2.3 Semantic tagging using rule-based approach

One of the more widely reported techniques for identification and classification of semantic information in medical/ biomedical domain is the rule-based or pattern-matching approach. For instance, (Long 2005) used UMLS (McCray et al. 1993) for identification and classification of semantic information (‘diagnoses’ and ‘procedures’) in discharge summaries. This method was based on analysing the structure of discharge summaries to locate required section headers (past medical history, discharge diagnoses) followed by identification of the required semantic information with the help of dictionaries and regular expressions. The identified semantic information was then coded by using a mapping of semantic entities ‘diseases’ and ‘procedures’ with their relevant UMLS semantic entities (Disease or Syndrome, Fungus, Injury or Poisoning,

Anatomical Abnormality, Congenital Abnormality, Acquired Abnormality, Mental or Behavioural Dysfunction, Hazardous or Poisonous Substance, Neoplastic Process, Pathologic Function). The corpus used in this research contained only 23 discharge summaries. Therefore, the applicability of this method on other types of structured and unstructured documents is likely to be very limited.

As mentioned in earlier section, the MetaMap system uses ontologies. In addition, this system also include rules for the semantic analysis of text (Aronson and Lang 2010). These rules split sentences in the form of phrases and associate identified concepts with semantic categories using ontologies (Aronson 2001). The MetaMap system was developed using MEDLINE abstracts, the structure of which is different from language used in clinical documents. In addition to this, the evaluation of the MetaMap system against any gold standard dataset was not reported. Therefore, the applicability of the system on other types of unstructured texts (such as medical narratives) was not claimed. The practical implementation and limitations of MetaMap on medical narratives are further discussed in Chapter 5.

Similarly, (Bashyam et al. 2007) also developed a module that extracted UMLS concepts from free text clinical radiology reports using a pattern-matching approach. They claimed that the processing speed of their module was faster than the MetaMap Transfer (MMTx) which is the Java version of MetaMap (Divita, Tse and Roth 2004).

MedLEE is another specialised NLP system that uses frame-based parser for analysis of grammatical structure in text. These grammatical structures then map to a frame and convert the frames into phrases. These phrases are then normalised to match with controlled vocabulary for encoding the concepts. This system was mainly developed to transform unstructured clinical narratives to structured and encoded text. The transformation of unstructured information varies from one type of report to another. Therefore, pre-processing for different reports with respect to their section headers was required (Friedman 2005). Since MedLEE, there has been a significant amount of research in lexicon-semantic mapping of various medical terminologies/controlled vocabularies to the UMLS and other terminologies (McCormick,

Elhadad and Stetson 2008). However, these systems failed to analyse long multiword phrases, as reported by (Sevenster, Ommering and Qian 2012).

(Skeppstedt, Kvist and Dalianis 2012) implemented the rule-based and terminology-based approach for the extraction of three semantic categories; 'Body Structure', 'Disorder' and 'Findings'. The main objective of this research was to evaluate the extent to which entities used in Swedish clinical notes are expressed in SNOMED CT. Their method was developed using SNOMED CT terminology because the translation of SNOMED CT was available in Swedish language. Moreover, these researchers used rule-based approach and lexical lookup using a combination of five different terminologies, and linguistic processing was done to refine the identification and classification of the semantic categories. The limitation of their approach is that they excluded the semantic category 'Qualifier Value' to be identified in these semantic categories. 'Qualifier value' such as 'Right', 'No', etc., indicates important information which cannot be passed on if excluded. By omitting 'Qualifier value' might effect the correct identification of other semantic categories such as 'Disorder' and 'Findings'. For instance, the concept 'No fever' should be categorised with the semantic category 'Findings'. This is due to the fact that the semantic category 'Findings' represents the results of clinical observation and 'No' represents the value in this concept. Therefore, excluding a 'Qualifier Value' will miss important information associated in this case. This also results in false analysis and may categorise 'Fever' as 'Findings' which in actual is 'Disorder'.

Another system used regular expressions for semantic analysis by analysing domain knowledge in physical notes that were annotated by two reviewers (Turchin et al. 2006). Their application identified semantic information related to 'blood pressure', with the blood pressure values and 'treatment' with the indication of medication in the text. Since this application was developed for this particular task and used data from only one source, it suffers from the limitation of applicability of regular expressions on any other text.

(Pakhomov, Buntrock and Duffy 2005) applied the set of rules on dictionaries including SNOMED CT, MeSH, RxNorm and Mayo Synonym Clusters (MSC). This was done for the

identification of ‘drugs’, ‘diagnoses’ and ‘signs and symptoms’ in clinical texts. Another information extraction system for semantic analysis was developed by (Liu et al. 2005). This system was based on GATE architecture and used a rule-based section filter, annotated information using subset of UMLS semantic categories, rule-based NegEx algorithm (Chapman et al. 2001) for negation detection and JAPE rules for the identification of specific attributes (Gleason score, tumour stage, status of lymph node metastasis) related to semantic information in pathology reports. Their approach used limited semantic categories for pathology reports.

On the basis of the literature review presented in this section, it was concluded that the existing rule-based or pattern-based systems cannot be successfully applied on texts other than those that were used in the actual development process. This is because these systems were developed and evaluated for a specific type of data and limited semantic categories. In contrast, it has been reported that systems based on machine learning approaches generally give better results in identification and classification of relevant semantic information in the medical domain. The more relevant machine learning systems and their limitations are discussed in the next section.

2.4 Semantic tagging using machine learning or statistical approaches

Recent applications in the medical/biomedical domain are mostly based on machine learning (ML) methods but ML approaches require large annotated corpora (training and test). This requirement is not only time consuming and expensive but also suffers with the limitation of access to large annotated data in the medical/biomedical domain (due to ethical issues). These points are highlighted in the studies summarised as follows.

(Sibanda et al. 2006) performed a semantic analysis of 48 discharge summaries. The semantic categories that were considered include ‘diseases’, ‘symptoms’, ‘treatments’, ‘tests’, ‘results’, ‘dosages’, ‘substances’ and ‘practitioners’. In this work, Link Grammar Parser (Sleator and Tamperley 1991) was used for the extraction of syntactic features, and support vector machines (SVMs) for training classifier. UMLS was used for mapping of the synonymous semantic categories. Their baseline system found the longest string that also included a head of noun

phrase in a noun phrase and then used UMLS to map relevant semantic categories. This baseline system was also compared with the MetaMap system and the results showed that MetaMap outperformed the baseline system only against one semantic category, 'disease'. The MetaMap system did not achieve better scores for other semantic categories. The low performance of the MetaMap was attributed to the fact that it used UMLS which did not contain noun phrases that occurred in the dataset used by these researchers. The results of these two systems (baseline and MetaMap) were then compared against their developed semantic category recogniser (SCR) which used multi-class support vector machines (SVMs). The performance of SCR was analysed using orthographic features (such as capitalisation, upper case, punctuation, etc.), lexical (such as bigrams, section headers), syntactic features (syntactic bigrams, head of noun phrase, part-of-speech), and ontological features (UMLS). The SCR outperformed the baseline system using all these features. However, on investigating a combination(s) of features, ontological features (UMLS) did not contribute in a better manner.

Another system was reported by (Taira and Soderland 1999) who used maximum entropy classifiers for semantic analysis and parsing structures in radiology reports. As in case of many other NLP systems, this system contained modules of a structural analyser, lexical analyser, parser and semantic analyser/interpreter. Structural analyser was a conversion from a rule-based system to a system that used a maximum entropy classifier. It structured sentences under section headers after analysing sections in the document (such as 'history', 'findings', etc.). The lexical analyser of this system used a medical lexicon for analysing semantic and syntactic features. It performed tokenisation of punctuations and normalisation of numeric values (such as dates, etc.). The parser and semantic analyser of their system were based on statistical methods; the parser formulated dependency structure arcs in a sentence which were then selected on the basis of high probability. On the other hand, the semantic analyser used the output of parser (arcs) and applied rules based on semantic features. The rules were then applied on unlabelled arcs to formulated logical relations which were then transformed into structured output frames. These frames contained attributes that identified the semantic categories of 'findings', 'therapeutic or

diagnostic procedure' or 'anatomic structure'. This system was evaluated using ten-fold cross validation via the use of a gold standard and achieved 89% precision and 90% recall. They also extracted UMLS concepts from radiology reports by using the vector space model (Bashyam and Taira 2005). The main limitation is the selection of a limited semantic categories and the use of specific type of data (radiology reports).

Another machine learning method was adopted by (Feng et al. 2008) who employed conditional random field (CRF) with active learning for semantic analysis of biomedical articles. This CRF model was specifically implemented to examine tract tracing experiments and used features based on lexical knowledge, surface words, context windows, window words and dependency features. These researchers also investigated different set of features in combination ('lexicon', 'lexicon + surface words', 'lexicon + surface words + window words', 'lexicon + surface words + window words + dependency features'). For all combinations, the system performed better than the baseline approach that just scanned words and phrases in the sentences from each lexicon. An overall F-score of 74% was reported on 16 documents. The limitation of this system was that the files contained variation in writing styles and thus needed more training data for better performance.

(Tang et al. 2013) used conditional random fields for the classification of three semantic categories 'Problem', 'Treatment' and 'Test' in discharge summaries. However, they investigated the use of structural support vector machines for the identification of concepts in discharge summaries. The identification and classification of these semantic categories were performed as a part of the global NLP challenge i2b2/VA 2010 (Uzuner et al. 2011). For this challenge, other teams also participated and the best performing system was by (Bruijn et al. 2010) who used the semi-supervised machine learning technique. Their system used the semi-Markov Hidden Markov Model (HMM) for identification of concepts in the corpus. Semi-Markov HMM was used to tag multi-token spans in the text (concept phrases) and the complete system achieved 85.3% f-measure.

In the biomedical domain, (Ananiadou et al. 2011) reported corpus annotation and approaches for the identification and classification of semantic categories in bacterial type IV secretion systems. These researchers presented four novel semantic categories for identification of gene and protein from the literature. They first developed training and evaluation corpus by using term extraction service to automatically identify multiword terms in the corpus. Two domain experts then reviewed negative examples in the corpus. After developing training and evaluation corpus, researchers evaluated three techniques listed below.

1. Dictionary-based approach [by matching longest term].

2. Dictionary-based approach with corpus enrichment [tagged terms found in training corpus were added to static dictionary and then matching was done].

3. Hybrid machine learning approach using a conditional random field with dictionary-based information.

F-measure score ranged between 18% to 96% for dictionary-based approach, 54% to 97% for dictionary-approach with corpus enrichment, and 68% to 93% for machine learning approach. This showed that the performance of the system was better in case of machine learning approach. However, this system was developed for a specific research task and data (biomedical text) and therefore, it cannot be used for other research questions targeting different data (medical narratives).

Other than above mentioned approaches, researchers also investigated and compared a different combination of approaches for the semantic analysis of clinical data. For instance, an NLP system named ‘HITEx - Health Information Text Extraction’ was developed in order to extract the key findings for airway diseases from 150 discharge summaries (Zeng et al. 2006). HITEx extracted semantic information that categorised principal diagnosis, co-morbidity, and smoking status. This system used UMLS concepts for semantic extraction of the principal diagnosis (Demner-Fushman, Chapman and McDonald 2009). HITEx has also used NLP components of GATE tool for specialised classification of semantic information. After basic language

processing modules and noun phrase chunking, this NLP system used UMLS concept mapper to match concepts in the text. To classify smoking status, the SVM classifier was used to extract single word features. Other semantic information such as principle diagnosis and co-morbidities were extracted using specific modules based on regular expressions.

In summary, the existing machine learning systems suffer from one or more of the following limitations. Failure at the complex level of synonymy, focus on any specific research question or corpus and the limited number of semantic categories using controlled vocabularies/ontologies. Thus, the conducted research did not provide flexibility to use data or annotations for general research purposes and the evaluation done by these researchers was restricted to specific research questions.

2.5 Summary

The identification and classification of semantic information in an ever increasing number of medical narratives in patient records is frequently required for several research applications such as statistical analysis, question-answering systems, negation detection, relationship extraction, etc. Different methods that are used for identification and classification of semantic information include ontology-based/dictionary-based approaches, rule-based or pattern-based approaches and machine learning or statistical approaches. On the basis of the review of literature presented in preceding sections of this chapter, we identified the following limitations and inadequacies of the existing approaches.

- Generalizability of methods for different datasets.
- Unavailability of (annotated) research data.
- Non-standard, inconsistent and limited semantic categories.

In addition, we noted that the problem of identification and classification of semantic information in medical narratives, including **concept phrases, concept abbreviations and complex multiword concepts**, has not been dealt together with in the existing literature. Besides helping in identifying the above mentioned limitations of existing system, the

background review also helped in the selection of appropriate techniques and resources for this research.

Considering the limitations of available systems and resources, this research focused on developing a generic and comprehensive rule-based semantic tagger, which was named SnoMedTagger. In the development of SnoMedTagger, I did not focus on mapping concepts with clinical codes present in the SNOMED CT healthcare terminology. However, the aim was to classify the concepts into globally known semantic categories that were derived from SNOMED CT. The proposed identification and classification technique is expected to facilitate consistent information exchange between domain users (such as medical/biomedical researchers) as well as between non-domain users (such as language researchers). Furthermore, the SnoMedTagger was designed to identify semantic information (such as paraphrases of concepts, complex multiword concepts, and abbreviations of concepts) in different datasets.

Chapter 3. The module for the anonymization of Protected Health Information

3.1 Introduction

Easy access to authentic data such as discharge summaries and progress notes is a major challenge for researchers. Only a few research datasets have been distributed as part of the shared Natural Language Processing – NLP research (*International Challenge: Classifying Clinical Free Text Using Natural Language Processing.* ; Pestian et al. 2007; *i2b2: Informatics for Integrating Biology & the Bedside* ; Uzuner, Luo and Szolovits 2007; Uzuner et al. 2008a). The organisers of these research tasks distribute datasets after the approval of specific data user agreements. Furthermore, these datasets contain annotations specific to research tasks designed for a challenge and cannot be readily used for other specific research tasks. For other research tasks, researchers face difficulties obtaining authentic annotated datasets despite their value for research.

A key reason behind the unavailability of real datasets for research is the need to respect the privacy of individuals such as patient, doctor, patient's relative etc. These real datasets cannot be made available to researchers without careful de-identification/anonymization of *Protected Health Information - PHI*. PHI is the information that can identify an individual. According to (Meystre et al. 2010), the terms de-identification and anonymization can be used interchangeably. The term de-identification refers to removing or hiding identifiers (PHI) from data while in anonymization, data is transformed to be completely anonymous. One difference is that in the case of de-identification it is possible to link data with identifiers while the anonymization process does not provide any link with identifiers.

In this study, we have followed the anonymization because the data needs to be distributed for research purposes and should not contain links to identifiers. Initially, we also came across the similar issue of data access for the development and evaluation of SnoMedTagger (Chapter 6) and this was addressed by participating in the fourth i2b2/VA shared NLP challenge (Hina et al.

2010). This corpus contained anonymized medical narratives and was found most feasible for the development and evaluation of the SnoMedtagger. Because, the SnoMedTagger was developed using the rule-based approach therefore required another evaluation dataset to prove the general applicability of the rules to more than one different dataset. Unlike the i2b2 corpus, the second dataset contained fictional information about individuals (Section 3.3 describes the origins of this dataset).

The second dataset was extracted by ResearchOne database (Crossfield and Clamp 2013a) from an EHR-Electronic Health Record system 'SystmOne'. SystmOne is a centralised clinical system (one EHR per patient) in the UK that provides the sharing of patient records between healthcare providers in the National Health Service (NHS). In addition to this, it contains functionality for integrated EHRs such as sending tasks, electronic prescribing, referral, appointment booking, bed management, etc. (Crossfield and Clamp 2013b). Real-data from SystmOne cannot be distributed for research purposes without the anonymization of PHI. For this reason a fictional dataset was created as an exercise using SystmOne to develop an anonymization module for SystmOne data. This data was representative of real-data and was referred to as 'Test dataset 1' in this research. The dataset was created during a training exercise for medical students to record a patient's consultation in SystmOne (explained in Section 3.3). Moreover, this dataset was novel and challenging for anonymization because it contained a mixture of natural language and clinical codes. The natural language elements of the dataset were known to contain references to named personal health information and identifiers of individuals and, as such, formed a rich training dataset for developing an anonymization module without using real patient's detail. The development of this anonymization module formed part of a project called 'e-Health GATEway to the Clouds'. This project aimed to establish a cloud-based research platform to support e-health records research. The project focussed on developing an anonymization module for the open source GATE tool (Cunningham et al. 2002) so that e-health records could be safely used by researchers following the best practice in ethics and governance (Smith et al. 2013).

The development of this anonymization module was separate from the major contribution of this thesis but considered as a valuable pre-processing step for using SnoMedTagger to anonymize PHI where it is a mixture of medical narratives and clinical codes.

To anonymize Test dataset 1, standard PHI categories provided by US Health Information Portability and Accountability Act 1996 (HIPAA) were investigated (Figure 3-2). According to the HIPAA guidelines, all proper names should be removed from the text or replaced by non-PHI. This can be done without the categorisation of their respective PHI categories (patient's name, place names, doctor's name, partner's name, nurse's name, etc.). But after consultation with the Health Sciences researchers, it was considered to be important to replace the names with their respective PHI categories (e.g., doctor's name, patient's name, etc.). This was done to maintain the readability of the text for analysis. For this purpose, the anonymization module was developed to anonymize data by replacing the identified PHI with their respective PHI categories, an approach described in more detail in (Hina et al. 2013). In this project, the anonymization of PHI in medical narratives is defined as a two step process.

1) Identification and classification of PHI.

2) Anonymization of PHI by replacing them with their respective PHI categories, shown in Figure 3-1.

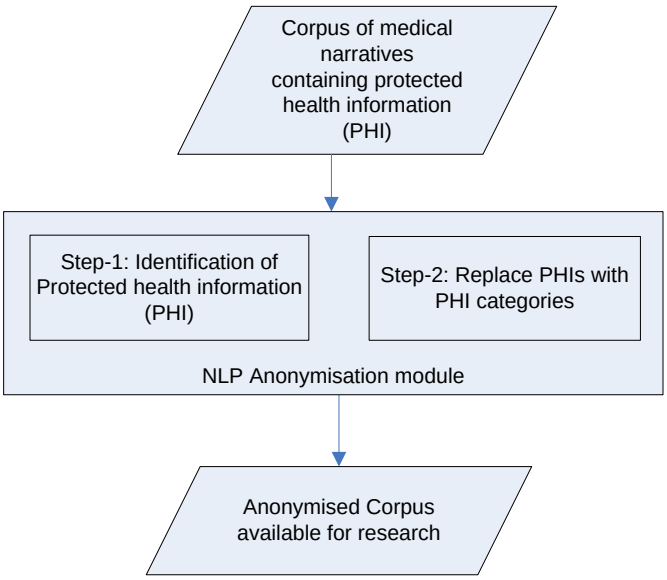


Figure 3-1: Steps in the anonymization process.

The rest of this chapter is organised as follows; Section 2 describes related work on automated anonymization/de-identification systems; the related work for this anonymization module is not included in 'Related work' (Chapter 2) of this thesis and is completely covered in this section. Section 3 covers the annotation of a gold standard corpus for the development and evaluation of this anonymization module; Section 4 contains the baseline method and method for the anonymization of PHI categories in Test dataset 1. The evaluation of the anonymization module is discussed in section 5, limitations are discussed in section 6 and lastly section 7 summarises the whole chapter.

3.2 Related work

Protection of information that identifies an individual should not be overlooked in the medical domain. In the United States, the Health Insurance Portability and Accountability Act provide 18 HIPAA categories (shown in Figure 3-2) for the de-identification of clinical data. This means, at least in theory, that after the removal of these 18 PHI categories, the data can be viewed as safe to use.

(Uzuner, Luo and Szolovits 2007) reported on a survey of anonymization tasks carried out as a part of the global Natural Language Processing (NLP) challenge, the organised by i2b2 project organisers. In this paper, the authors described the process of annotating the gold standard for the de-identification challenge. The i2b2 challenge organisers prepared data by annotating protected health information (PHI) and replacing PHI with realistic surrogates for evaluation. The gold standard data was compiled for the de-identification of following eight categories; Patient, Doctors, Hospitals, IDs, Dates, Locations, Phone numbers, Ages. This gold standard was first annotated by an automatic system and then validation was manually done by three annotators. After validation, annotated PHI was replaced by realistic surrogates. Inter-annotator agreement was not reported by these authors.

Other than machine learning approach, some researchers proposed methods of using 'dictionaries' and 'natural language processing using features and heuristics' for the anonymization of medical records (Tveit et al. 2004). Their methods were proposed for

anonymization of patient records but mainly focused on the anonymization of general practitioner records. They used a Norwegian corpus which was challenging because its linguistic features varied from the English language and existing approaches could not be used. In the first step, they constructed dictionaries using their own corpus and some external dictionaries from other sources such as dictionaries of medical names, geographical names and Norwegian person names etc. In their second step, all dictionaries were compiled in a single dictionary to perform the exact matching of names. Moreover, a suffix tree was used to improve the matching performance and the matched names were tagged with their respective types. Non textual types (such as dates, phone numbers, security numbers, etc.) were identified using the suffix tree, then all tagged words which had multiple types were investigated and untagged words were manually reviewed by a local clinician for tagging. Finally, all tagged words were replaced by pseudonyms. These researchers have not shown or discussed any aspect of validation or evaluation of their work which shows the limitation of their approach on other datasets.

Another de-identification program was reported by (Marciniak, Mykowiecka and Rychlik 2010). They developed a rule-based system to anonymize patient's personal information. The method was based on the identification of a patient by their surname, forename and date of birth. This approach might fit to the structured patient's records in which each document contains a surname, forename and date of birth but will not work for unstructured documents that contain random clues about someone’s personal information. These authors reported a number of documents in their evaluation and discussed the problems encountered during the anonymization of data but this method did not guarantee its general applicability on unstructured text. One distinct and useful implementation in this research was the creation of key code for each patient so that the patient's record could be reused in the future.

This space is deliberately left blank due to pagination.

1. Names [fictitious names can be used to facilitate writing]
2. Geographic locations smaller than a state/county, including zip codes or post codes
3. All elements of dates except years relating to individuals
4. Telephone numbers
5. Fax numbers
6. Email addresses
7. Social security numbers
8. Medical records numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate and license numbers
12. Vehicle identifiers
13. Device identifiers and serial numbers
14. Universal resource locators (URL)
15. Internet protocol addresses (IP addresses)
16. Biometric identifiers
17. Full face photographs
18. Any other unique identifying number, characteristic or code

Figure 3-2: HIPAA 'Safe Harbour' categories[4].

In contrast with the above mentioned systems, (Szarvas, Farkas and Kocsor 2006) presented a de-identification method which was presented in the first i2b2 global NLP challenge on clinical data. They reported a novel iterative machine learning approach for named entity recognition (NER) using semi-structured documents. This method first tags all entities which were present in structured parts of the document and then this information was further used to find other PHI in unstructured parts of the text. To find PHI, these researchers employed orthographical features, frequencies of tokens, PHI phrases and lookups (dictionaries of locations names, diseases, non PHI tokens, etc.) for word-level classification. Using this feature set, a combination of two machine learning classifiers (Boosting, C4.5) was trained in three phases and successfully achieved 99.7534% of f-measure on the evaluation set. This method is

4 http://www.press.jhu.edu/journals/narrative_inquiry_in_bioethics/HIPAA_Safeharbor.pdf

specifically developed using semi-structured documents, therefore not ensuring its general applicability on other datasets.

(Ruch 2000) also reported an anonymization system to de-identify name, address, phone number and date of birth and noted the problem of ambiguity in PHI identifiers. For instance, 'River' can be written as common noun as well as proper noun; ‘River Song’ is a character in a BBC-TV programme. Such nouns do not always have clues to help resolve the ambiguity so should be removed/ replaced if present as PHI in the corpus. Therefore, to tackle these problems, a de-identification system was developed on more than 40 rules. The corpus used in this research was a mix of the German and English languages and was split into two sets;

1) 20% of the corpus was used to set up the system.

2) 80% of the corpus for evaluation.

This system was based on;

1) MEDTAG lexicon for lexical resources.

2) Rule-based morphosyntactic (MS) and a word sense (WS) tagger for the disambiguation task.

Although a 99% success rate was reported in this paper, their method was still not suitable for the dataset (Test dataset 1) used in this research because of different PHI categories.

The majority of the work was done on structured data. However, some researchers also worked on both structured and unstructured data. For instance, (Gardner and Xiong 2009) developed a conceptual framework named HIDE (Health Information DE-identification) for de-identification of PHI in both structured and unstructured data. They employed a Bayesian classifier, sampling based techniques and conditional random fields based techniques for the extraction and identification of sensitive information from the data. Their method also provided the benefit of data linkage by using an identifier for an individual record and also provided three flexible options for the de-identification; full de-identification, partial de-identification and statistical de-identification. Preliminary results showed overall accuracy of 75%-98% for the de-identification of name (with respect to how long they extend), age, account number, medical record number

and date. (Uzuner et al. 2008b) used different artificial and real-time corpora containing personal health information for a comprehensive analysis and presented a de-identifier named Stat De-id, based on the support vector machine (SVM) method and the local context. The approach was successful in proving that Stat De-id using SVM and local context outperformed over four systems; 1) SNoW (Roth and Yih 2002), 2) IdentiFinder (Bikel, Schwartz and Weischedel 1999), 3) Dictionaries + Heuristics and 4) Conditional random fields (CRF). Stat De-id de-identified the following seven PHI categories in discharge summaries; Patients, Doctors, Hospitals, IDs, Dates, Locations, Phone numbers. This method used a large number of features (syntactic, syntactic bigrams and semantic features). The limitation of the Stat De-id system was reported in terms of the absence of local context in sentences. In the i2b2 de-identification challenge mentioned above, another participant team (Aramaki et al. 2006) learned local, global and external features by using conditional random fields - CRF. They used Beginning-Inside-Outside (BIO) tagging to identify chunks in tokens. External features used in this work included dictionaries of people, locations and dates; global features included sentential features to mark sentences and tokens from the previous sentence.

A recent review was done by (Meystre et al. 2010) on automatic de-identification of systems developed after 1995. This review helped me in the completion of a literature review on automatic de-identification systems/tools. According to this review, the majority of work was done on structured data and very few researchers have focused on narratives. The review concluded that de-identification systems mainly address the common PHI category of names but also consider other, different PHI categories. Having different PHI categories is one of the reasons why one de-identification system cannot easily be compared with other de-identification systems. They analysed 18 systems including some discussed earlier in this section (Ruch 2000; Aramaki et al. 2006; Szarvas, Farkas and Busa-Fekete 2007; Uzuner, Luo and Szolovits 2007; Gardner and Xiong 2008; Uzuner et al. 2008b) and a further 12 which will be discussed in the following text. All 18 systems analysed in this review de-identify the general categories of names, ages, dates, contact details, hospitals and healthcare providers, locations and ids. In

terms of methods, mainly pattern-matching algorithms and machine learning methods were adopted while some systems used both approaches.

(Beckwith 2006) developed an open source tool for the de-identification of pathology reports. The system was named HMS (Harvard Medical School) Scrubber and followed three steps to complete the process of de-identification. In the first step, all pathology reports were converted into an XML format to separate headers and text. This step gives a structured look by separating important PHI into headers such as date of birth, medical record number, social security number, accession number and pathology department. Then in the second step, pattern-matching was performed using regular expressions to find patterns of date, telephone number, etc. In the last step, string matching was done to identify and remove person names and location names. HMS Scrubber achieved 98% of recall on 1800 reports. This system was meant to process structured reports and contained PHI categories different from those used in this research.

In another study, researchers have also used rules, lookup tables and regular expressions to de-identify PHI in medical documents (Gupta, Saul and Gilbertson 2004; Neamatullah 2008). The PHI categories used by these researchers were specific to their individual datasets and cannot be compared against different datasets. Similar to the system developed by (Beckwith 2006) and (Gupta, Saul and Gilbertson 2004), another system named MeDS was reported by (Friedlin and McDonald 2008) which used regular expressions, headers and dictionaries (for persons and locations). They used around 50 regular expressions to identify and remove misspelled names in the corpus. MeDS was evaluated on two different datasets.

1) 2400 reports (laboratory reports, narrative reports, mixed source reports).

2) 1193 surgical reports.

On the first dataset, MeDS was able to de-identify 99.06% of the HIPAA identifiers and 98.26% of the non-HIPAA identifiers. On second dataset, MeDS identified 99.47% of the HIPAA identifiers and 96.23% of the non-HIPAA identifiers.

A system called concept-match scrubber was developed by (Berman 2003). Initially, the documents were pre-processed by parsing text into words, sentences and stop words, and then the open source nomenclature UMLS was used to match and replace standard terms with their respective terms and codes. Then, all non-matching terms were replaced by a blocking tag. This system might help researchers working on statistical analysis but might not be helpful for the contextual analysis of documents. The authors of this work did not report any actual standard measurements of precision and recall but they expected their system to achieve high recall because documents only retained identifiers containing stop words.

The review also included a rule-based de-identification system named 'Scrub' which was developed by (Sweeney 1996) who used several parallel detection algorithms and local dictionaries to de-identify proper names (first names, last names, full names), addresses, states, countries and cities. In this study, two set of experiments were conducted. In the first experiment, humans were employed to identify PHI in letters written by physicians. The second experiment was a computer-based approach that used a detection algorithm and knowledge sources. There was a separate detection algorithm for each entity (PHI) and the algorithm reporting the highest value of likelihood was considered in case of ambiguity. The Scrub system successfully de-identified personally identifying information (up to 99%-100%) in comparison with database lookup (achieved 32%-37%) and database lookup with cues (32%-84%).

In comparison with the well known rule-based and pattern-matching systems, a different approach was adopted by (Morrison 2009) who used the natural language processing (NLP) system MedLEE to identify and extract medical concepts in reports. As a result of this extraction, the corpus only contained medical concepts without PHI. The output MedLEE was reviewed by a physician and only 3.2% of PHI were detected in the corpus. This approach suffers from the limitation of maintaining contextual information in the corpus.

A technique based on a lexicon of names and UMLS was presented by (Thomas et al. 2002) which used an augmented search and replace algorithm to identify proper names in the corpus. Their method also included the use of regular expressions to identify prefixes and suffixes of

names associated with proper names. This system was evaluated against a manually developed gold standard of 1001 pathology reports and identified 98.7% of names in the narrative section and 92.7% of names in the whole corpus.

Other than the use of dictionaries and pattern-matching using regular expressions, some researchers adopted machine learning and statistical approaches for the task of de-identification in medical records (Taira, Bui and Kangarloo 2002; Guo 2006; Hara 2006; Wellner 2007; Szarvas, Farkas and Busa-Fekete 2007; Uzuner et al. 2008b). Among these (Taira, Bui and Kangarloo 2002) used statistical modelling to identify names in patient reports while (Wellner 2007) used two toolkits Lingpipe (*LingPipe 4.1.0.*) and Carafe[5] for the identification of named entities in the corpus.

In the machine learning approaches, (Guo 2006) used the support vector machine (SVM) method, and the well-known named entity recognition system, ANNIE (Cunningham et al. 2002). ANNIE was used to pre-annotate the training set with a person's name, date, etc. Then multiple features were used to train the SVM machine's learning classifier including date features, doctor name features, etc. This system participated in the first i2b2 NLP challenge and achieved precision, recall and f-measure greater than 86%. (Hara 2006) also used SVM to develop a de-identification system. In their system, SVM was used to perform named entity recognition (NER) in medical reports. This system also participated in the i2b2 challenge of de-identification and achieved 92% (approximately) of f-measure. Their method first used pattern matching to identify section headers, then regular expressions were used to identify dates and phone numbers. A sentence classifier was also used to identify PHI in sentences and finally an SVM based text chunker was used to identify location, patient, age, etc.

As reviewed in this related work, researchers developed de-identification/anonymization systems for different named entities specific to the requirement of their datasets. These systems were developed using different methods for both structured and unstructured datasets which cannot be compared directly since they have used different named entities (PHI categories) and

[5] http://carafe.sourceforge.net/

are based on different datasets with a different nature of data (structured / unstructured). In comparison, the Test dataset 1 used in this research was completely different from the data used by previous researchers. This is because it was drawn directly from an EHR record of a consultation rather than from a natural language report and, as such contained a mixture of natural language and clinical codes. In addition to this, it contained four different PHI categories (Patient Names, Doctor Names, Other Names, and Place Names) which were not investigated together in the work done by previous researchers. The PHI category 'Other Names' is a new category for all names other than patient's names and doctor's names.

As mentioned in Section 3.1, this Test dataset 1 was required for the evaluation of the semantic tagger (SnoMedTagger) developed in this research. Thus, there was a need for an anonymization module for SnoMedTagger in case of data containing identifiers. For this reason, 18 HIPAA PHI categories were investigated and the above mentioned four PHI categories were customised according to HIPAA rules for the Test dataset 1.

3.3 Gold standard corpus for development and evaluation of anonymization module

The Test dataset 1 used in this research was created as a result of a large number of teaching lab sessions conducted for medical students. A paper-based form containing patient's protected health information (PHI) details was given to each student and a recorded consultation video was shown to the medical students. The medical students were then asked to record this consultation in SystmOne. Students were free to record the consultation in the form of a mixture of clinical codes (READ codes – *coded vocabulary for clinical terms*) and narrative text. Because the data was created as part of a teaching exercise, the data contained fictional names of patients. Most of students used natural language instead of clinical codes to record their observations although some of them used both clinical codes and natural language and some used almost exclusively codes. The data was challenging for the anonymization module because alphanumeric clinical codes were written within a natural language free text consultation record. This sample data is shown in Figure 3-3.

Because of the unavailability of large annotated data, a rule-based approach was adopted in the development of this module. To develop a rule-based anonymization module, 15% of Test dataset 1 was divided into a ‘Development set’ and the rest 85% ‘Evaluation set’ was left for the evaluation of the anonymization module.

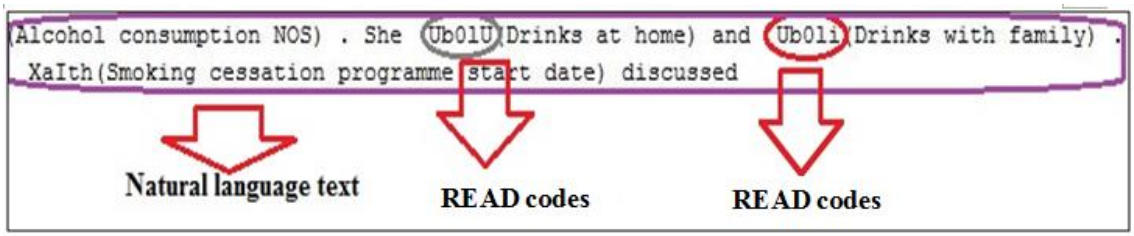


Figure 3-3: Sample text from corpus containing natural language and READ codes.

In order to prepare a gold standard corpus for the development and evaluation of the anonymization module, an annotator manually reviewed Test dataset 1 to identify possible PHI categories present in the corpus. The corpus contained the following four PHI categories.

1. Patients Name

2. Doctors Name

3. Other Name (person names other than patient’s names and doctor’s names)

4. Place Name

There were few person names in the corpus which were not under the PHI categories of 'Patients Name' and 'Doctors Name'. Therefore, all these names were categorised under the category of 'Other Name'. As mentioned earlier in Section 3.1, after consultation with Health Science researchers the identification and classification of names with respect to their roles/occupations was decided as a requirement of this project. The significance of this classification was to maintain the readability and analysis of the text for researchers. All PHI identified by the annotator were then manually annotated with their respective PHI category to produce a gold standard corpus using an open source annotation tool GATE (Cunningham et al. 2011). Corpus measurements and gold standard annotations are tabulated in Table 3-1.

Table 3-1: Corpus measurements and gold standard annotations.

Types	Development set (15%)	Evaluation set (85%)	Total
Corpus measurements			
Patient Records	301	1683	1984
Tokens	23031	167065	190098
Sentences	1298	9889	11187
Gold standard annotations			
Patients Name	376	2117	2493
Doctors Name	1	6	7
Other Name	2	5	7
Place Name	2	25	27

3.4 Anonymization of protected health information in medical narratives

The anonymization module was developed for the anonymization of four protected health information (PHI) categories in the corpus containing medical narratives and clinical codes. In the first step of anonymization, identification and classification of PHI were required. For this reason, an existing named entity recogniser, 'A nearly new information extraction - ANNIE', provided in the open source GATE tool (Cunningham et al. 2002) was modified as a baseline system (explained in next section); then on the basis of the limitations observed in the baseline system, a rule-based system was developed for the identification and classification of PHI (described in Section 3.4.1). The reason behind implementing a rule-based approach was the unavailability of a large annotated corpus (training and test) which would be required for machine learning methods.

3.4.1 Modification of existing named entity recogniser as baseline system

In Natural Language Processing (NLP) applications, the development usually starts with a basic and simple approach and then progresses to an advanced application. This basic approach is called the ‘baseline’ application/approach. The results produced from this application, the baseline results, are used for comparison throughout the application development.

For the identification and classification of PHI in patient records, we investigated the use of ANNIE. This system only identifies and classifies proper names and does not distinguish names corresponding to their role/ occupation ('Patients Name', 'Doctors Name' and 'Other Name'), which was the requirement of this project. Therefore, we modified the ANNIE application as baseline system by adding simple rules on the dictionaries/gazetteers of person's names, titles, and by adding a dictionary of 'others' which contained roles/occupations other than the doctor's occupation (For example, 'sister', 'partner', 'nurse', etc.).

The application pipeline of our baseline system included tokenisation of the corpus, splitting sentences, dictionaries/gazetteers, tagging the corpus with part-of-speech tags and Java Annotation Pattern Engine - JAPE transducers for the development of rules (Cunningham, Mayard and Tablan 2000), as shown in Figure 3-4. The general syntax of JAPE grammar rule is;

Rule: Rule Name {Pattern} --> Rule {Action}

The left hand side of each rule contains a pattern which is meant to perform the right hand side action subject to match the rule. The JAPE rules were applied independently without using output of other rules. A combination of rules creates a phase and a number of phases combine to form a grammar.

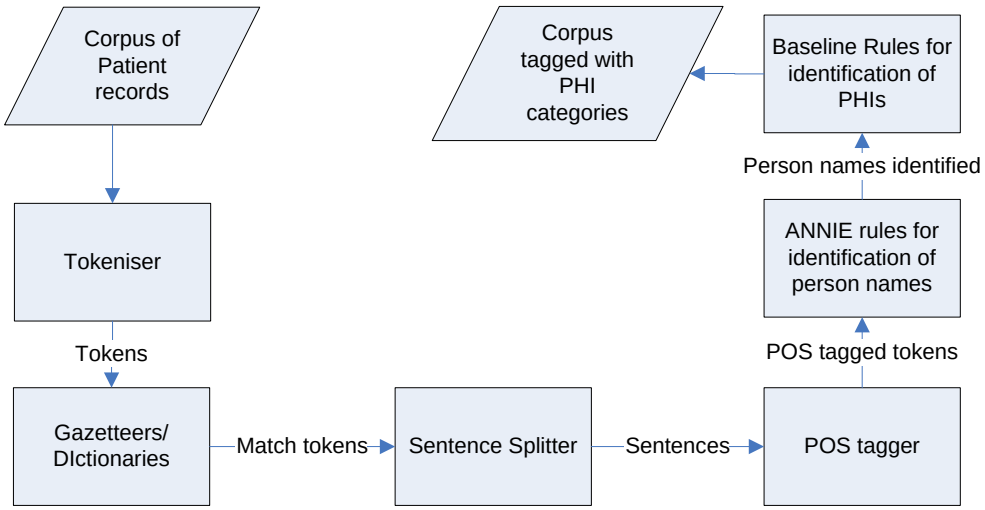


Figure 3-4: System flow of baseline system for anonymization.

In the first step, the ANNIE tokeniser was used to tokenise the corpus and then modified ANNIE gazetteers were used to match tokens for the identification of PHI. These ANNIE gazetteers are plain text files which contain a list of entities entered per line. Each plain text file represents a separate set of named entities (such as 'names', 'locations', etc.). After the identification of names matched by the dictionaries/gazetteers, the ANNIE sentence splitter was used to split sentences. This sentence splitter was required prior to the ANNIE semantic tagger; then the ANNIE part-of-speech (POS) tagger was used in the pipeline to assign POS tags to tokens. This POS-tagging was also required for the application of the ANNIE semantic tagger because the semantic tagger contained rule-patterns based on POS tags.

After applying the basic language processing modules from ANNIE, a set of hand-crafted rules was added to the application pipeline for the identification of the four PHI categories; 1) Patient Names, 2) Doctor Names, 3) Other Name and 4) Place Name. The baseline system mainly used dictionaries to identify these PHI categories and in addition to dictionaries it applied ANNIE rules for the identification of proper names of persons and simple JAPE rules for differentiating between ‘Patient Names’, ‘Doctor Names’, ‘Other Name’ and ‘Place Name’.

For instance, for the identification of ‘Patient Names’, rules for the classification of person names were used from the ANNIE application; ‘Doctor Names’ were identified by rules searching ‘titles’ matched from the dictionary before proper names. Similarly, ‘Other Names’ were identified by rules that searched clues matched by dictionary of ‘others’ which appeared before proper names (Other Names). Finally, the ‘Place Names’ were identified by the ANNIE dictionaries of ‘country’, ‘cities’, ‘company’, ‘department’, etc.

This baseline system was evaluated using standard information extraction metrics; Recall, Precision and F-measure (Sokolova and Lapalme 2009). The formulas of recall, precision and f-measure are provided in Section 3.5. As mentioned earlier, this baseline system was developed using 15% of the Test Dataset 1 (Development set) and evaluated on 85% of the Test dataset 1 (Evaluation set). The baseline system achieved overall 75% of f-measure on the Development

set and 76% f-measure on the Evaluation set; the detail measurements on each category for the development set is given in Table 3-2 and for the evaluation set is given in Table 3-3.

Table 3-2: Evaluation metrics of baseline system on the Development set.

Baseline results on Development set								
PHI Categories	**Correct matches**	**Partial Matches**	**True positives (Tp)=correct matches + partial matches**	**False Negatives (Fn)**	**False Positives (Fp)**	**Recall(%)**	**Precision(%)**	**F-measure(%)**
Patients Name	41	277	318	58	135	85	70	77
Doctors Name	0	0	0	1	0	0	0	0
Other Name	0	0	0	2	0	0	0	0
Place Name	2	0	2	0	18	100	10	18
Micro average	43	277	320	61	153	84	68	75

Table 3-3: Evaluation metrics of baseline system on the Evaluation set.

Baseline results on Evaluation set								
PHI Categories	**Correct matches**	**Partial Matches**	**True positives (Tp)=correct matches + partial matches**	**False Negatives (Fn)**	**False Positives (Fp)**	**Recall(%)**	**Precision(%)**	**F-measure(%)**
Patients Name	258	1524	1782	335	622	84	74	79
Doctors Name	1	0	1	5	0	17	100	29
Other Name	0	0	0	5	0	0	0	0
Place Name	24	0	24	1	147	96	14	24
Micro average	283	1524	1806	346	769	84	70	76

In this evaluation, ‘Partial matches’ were only counted for the cases where the system was unable to identify ‘Correct matches’ (full names in the context of this study). For instance, ‘Partial match’ will not be counted if the full name ‘John Smith’ is correctly identified.

From Table 3-2 and Table 3-3, it can be observed that in the case of ‘Patient Names’, dictionaries were able to identify individual names (Partial names) but were not able to identify all full names (Correct matches of full names). Other PHI categories also suffered from low performance measurements due to insufficient clues for the identification and classification of PHI.

3.4.2 Rule-based module for identification, classification and anonymization of PHI

After the implementation of the baseline system, it was observed that the dictionaries not only lose the required information but also identified the false information. Therefore, more rules were developed to resolve the following highlighted issues (identified by the baseline system) observed in the Development set;

1. There were some ***Asian names*** and ***Nicknames*** in the corpus, which were not present in the dictionary.

2. Some names in the corpus were not written in the proper ***format***.

For example 'Davina TRN Smith' is a patient name which is not in the dictionary but 'Davina' and 'Smith' were in the dictionary. In this case, 'TRN' can be assumed as a set of ***initials*** but was not reflecting the ***initials*** of this name. Another patient name 'mrs. parsons' was missed by dictionary application because it was written in ***lower case letters***.

3. ***Coded information*** in the corpus was picked as the short form of the place names and patient names. For example, in READ code 'Xa0NZ', 'NZ' was picked up as the short form of the New Zealand which was in the dictionary.

4. Clinicians can write either a patient's full name or first name or surname in medical narratives. Names such as 'May', 'Little', 'Short', 'Long', etc., can also occur in the form of ***verbs*** and ***adjectives*** in medical narratives. Thus, there is a chance of identification of such verbs/adjectives as names. A relevant example is shown in Figure 3-5.

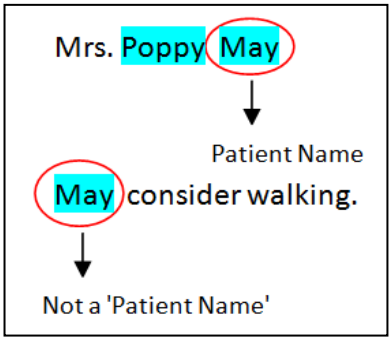


Figure 3-5: Example of the output of the baseline system.

5. The ***medical terms*** can be determined as proper names in the medical narratives for instance, 'Ray' was identified as proper name in 'X Ray'. Another example was 'TIA' which is an abbreviation of the medical term ' transient ischemic attack' was determined as a proper name in the corpus.

Because of problems described above, the existing general named entity recognition systems appear inappropriate for medical narratives.

Similar to the baseline system, the dictionaries of locations, person_first (male) and person_first (female) names from the ANNIE application were used in the development of the rule-based anonymization module. The corpus was a tab delimited file, and therefore all the nicknames and Asian names were extracted by exporting corpus into a Excel spreadsheet. All extracted names were then added to the dictionary of names.

The dictionary of location names was used to identify 'Place Names' and the remaining two dictionaries of names (person_first (male) and person_first (female)) were integrated in to a single dictionary of names. This single dictionary was compiled because the category of 'Patient Names' and 'Doctor Names' did not require categorisation of male and female names. Therefore, in addition to the names in dictionaries, a rule-based anonymization module was developed by analysing problems identified in the development corpus (explained in the next section). The complete system flow of anonymization module is shown in Figure 3-6.

The pre-processing of this anonymization module included basic language processing steps (tokenisation, split sentences, part of speech tagging). After applying basic language processing modules, the dictionaries were added in to the application pipeline to look up names in the corpus. Finally, the rule-patterns were added to identify categories of 'Patient Name', 'Doctor Name', 'Other Names' and 'Place Names'.

The dictionary named 'others' was added to identify names other than patient names and doctor names. This dictionary included roles and occupations which do not represent any patient name or doctor name. For instance, 'Nurse', 'Nurse practitioner', 'brother', 'partner', etc. can represent

person names in the corpus. The reason for separating dictionary of the 'others' was to distinguish the rules for identification of all names other than patient names and doctor names.

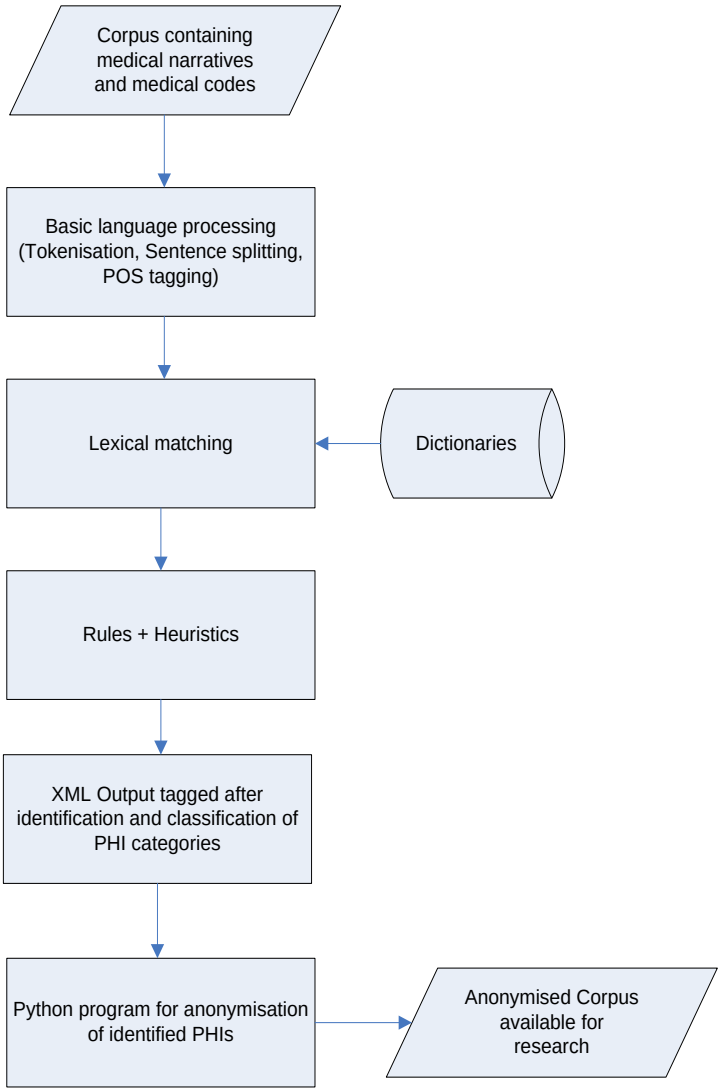


Figure 3-6: System flow of anonymization module.

Another dictionary of 'Noplace' was compiled containing the terms which were wrongly identified by dictionary application and for which general rules were not applicable. For all such cases, the rules were developed to restrict the false positives of ‘Place Names’. Some example rules and false positives are shown in Table 3-4. In addition to these dictionaries, a rule-based approach was adapted for the identification of PHI categories.

The next section describes the development of rules for identification and anonymization of four PHI categories in medical narratives. The anonymization of protected health information (PHI) was dealt with in two steps.

1. Identification of PHI and classification with their respective PHI categories, shown in Figure 3-7.
2. Anonymization of PHI by replacing them with their respective PHI categories, shown in Figure 3-8.

Table 3-4: Examples rules to restrict the false positives of patient names and place names.

Rules to restrict false positives of 'Patients Name' and 'Place Name'	**False positives identified by dictionary application**

Rule: Nopatient
(
(
{Token.orth==number}
(NAME)
)
):match
-->
:match{
inputAS.remove(bindings.get("match").iterator().next());
}
	X76Li is READ code in which 'Li' was identified as 'Patient Name'.

Rule: Noplace
(
{Lookup.majorType==noplace}
):match
-->
:match{
inputAS.remove(bindings.get("match").iterator().next());
}
	1. Seemed Nice.
	2. NICE guidelines
	3. Split up with husband.
	Blue highlighted terms were wrongly identified as 'Place Names'

This space is deliberately left blank due to pagination.

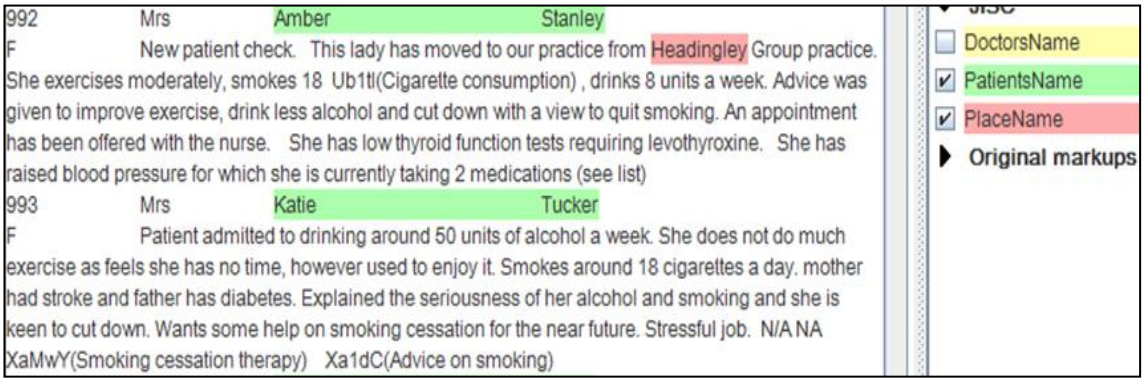


Figure 3-7: Identification and classification of PHI categories.

```
992 Mrs patient name            F   New patient check.   This lady has moved to our practice from place name Group practice.
She exercises moderately, smokes 18  Ub1tI(Cigarette consumption) , drinks 8 units a week. Advice was given to improve exercise,
drink less alcohol and cut down with a view to quit smoking. An appointment has been offered with the nurse.
She has low thyroid function tests requiring levothyroxine.   She has raised blood pressure for which she is currently taking
2 medications (see list)

993 Mrs patient name            F   Patient admitted to drinking around 50 units of alcohol a week. She does not do much
exercise as feels she has no time, however used to enjoy it. Smokes around 18 cigarettes a day. mother had stroke and
father has diabetes. Explained the seriousness of her alcohol and smoking and she is keen to cut down. Wants some help on
smoking cessation for the near future. Stressful job.  N/A NA XaMwY(Smoking cessation therapy)    Xa1dC(Advice on smoking)
```

Figure 3-8: Anonymization of PHI categories.

3.4.2.1 Identification and classification of PHI with their respective PHI categories

To develop useful rule-patterns of proper names for the identification of PHI and their classification with PHI categories, first a baseline system (explained in Section 3.4.1) was tested on the development set which used the dictionary to match names in the corpus. The observations showed that many false positives were marked by the baseline system (Table 3-5). As mentioned in section 3.4.2, the dictionary of names was compiled by extracting all names and nicknames from the corpus. The false positives such as ‘am’ or ‘read’ identified by the baseline system were due to nick names present in the corpus.

Table 3-5: Examples of false positives identified by dictionary application and rule-patterns developed to restrict them.

Rule-patterns	**False Positives**
Token.category!=VBP	Not drinking in '**am**'.
Token.category!=VBN	She has '**read**' over the info regarding smoking.
Token.category!=VB	...cut '**short**' the consumption of alcohol.
Token.category!=RB	She feels a '**little**' guilty for drinking.
Token.category!=MD	...that diabetes '**may**' be related to cough.
Token.category!=JJ	'**green**' spit
Token.category!=lowercase	'**little**' or no exercise.

In addition to the dictionary matching, the correct identification of individual names needed a number of rule-patterns. Therefore, a general 'Macro' rule was developed to identify individual names irrespective of their relevance with a PHI category, shown in Figure 3-9.

```
Macro: NAME
(
 {Lookup.majorType == Names,
 Token.category!=VBP,
 Token.category!=VBN,
 Token.category!=VB,
 Token.category!=RB,
 Token.category!=MD,
 Token.category!=JJ,
 Token.orth!=lowercase}
)
```

Figure 3-9: Macro rule for the identification of proper names.

First all single word proper names of patients were filtered by the 'Macro' rule by token matching in the dictionary. Then false positives were restricted using rule-patterns that checked the category features of the tokens. Examples of false positives are shown in Table 3-5 with the patterns developed for restricting them. This Macro rule identified single names (First name/ Middle name/Last name) in the corpus as these single word names appeared in the natural language free text. This general Macro rule was then used in the development of rule-patterns for names under specific PHI categories of 'Patient Names', 'Doctor Names' and 'Other Names'.

3.4.2.1.1 Identification and classification of 'Patients Name'

In the development set of medical narratives, it was observed that the medics used different formats to write proper names. For instance, some medics used the first name or full name while others used the initials of middle names in full names. Similarly, a range of formats were observed in the development corpus to develop useful rules for the identification and classification of patient names. In addition to single word names, each consultation note started with full names along with nicknames. These names were separated with a tab space at the start of each consultation note and appeared within the text in a range of different formats. Some example formats of names are given in Figure 3-10 which were identified by the macro rule.

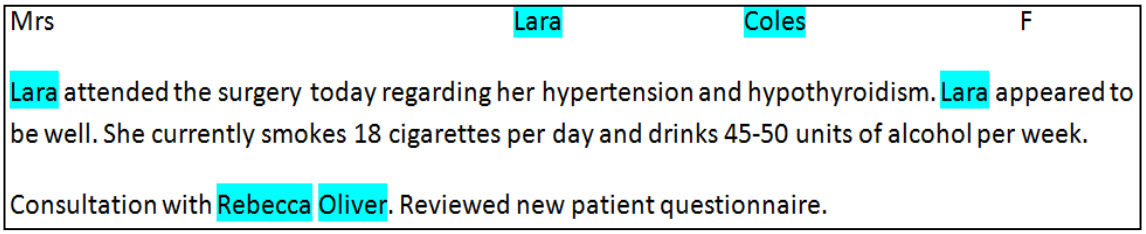


Figure 3-10: Output of Macro rule.

For the identification of full names of patients, several rule-patterns were developed by analysing the Development set. These rule-patterns were able to identify full names including initials and nicknames. Some example JAPE rules which successfully identified patient names are presented in Figure 3-11.

This space is deliberately left blank due to pagination.

```
Rule: Patient Name
(
        (
        (NAME)                                  //Macro Rule for identification of names
        (SPACE)                                 //Macro Rule for capturing spaces and tab spaces
        (NAME)
        )                                       // identify full names such as 'John Smith'
        |                                       //OR
        (
        (NAME)
        {Token.kind==punctuation}
        (NAME)
        )                                       //identify names such as 'Smith,John
        |                                       //OR
        (
        (NAME)
        (SPACE)
        {Token.orth == upperInitial, Token.length =="1"}
         (SPACE)
        (NAME)
        )                                       //identify names with initials, for example 'John A Smith'
): label
-->
: label.PatientsName= {Rule=Patient Name}
```

Figure 3-11: Example rules for identification and classification of Patients Name.

3.4.2.1.2 Identification and classification of 'Doctors Name'

During the analysis of the Development set, it was also observed that the names of doctors were written along with their titles or occupations (Dr, GP, Doctor, etc.). Therefore, a separate dictionary storing titles of doctor was created. These titles were used as a clue to identify names of Doctors/ General practitioners in the corpus. The same Macro rule (Figure 3-9) for names was reused to develop rules for the identification of doctor names. The rules for the identification of doctor names were developed by applying simple heuristics using title clues as shown in Figure 3-12. After the identification of 'Patients Name' and 'Doctors Name', rules were developed to identify 'Other Name' (explained in the next section).

This space is deliberately left blank due to pagination.

```
Rule: Doctor Name
(
        (
        {Lookup.minorType == title}                    // Dictionary of doctor titles
        (SPACE)
        (NAME)                                         //identify doctor name such as 'Dr Fred'
        )
        |                                              // Or
        (
        {Lookup.minorType == title}
        {Token.kind==punctuation}
        (NAME)
        )                                              //identify doctor name such as 'Dr. Fred'
        |                                              // Or
        (
        {Lookup.minorType == title}
        {Token.kind==punctuation}
        (NAME)
        {Token.kind==punctuation}
        (NAME)
        )                                              //identify doctor name such as 'Dr. Fred,Smith'
        |                                              //Or
        (
        ({Lookup.majorType == title}
        {Token.orth == upperInitial, Token.length =="1"}
        ({Token. string == "."})?
        )+
        (NAME)
        )                                              ////identify doctor name such as 'Dr A.B.Smith'
): label
-->
: label.DoctorsName= {Rule=Doctor Name}
```

Figure 3-12: Example rules for the identification and classification of Doctors Name.

3.4.2.1.3 Identification and classification of 'Other Name'

The medical narratives used in this research contained a few names other than patients and doctors. These names were related to other roles/occupations such as nurse, partner, husband, etc. These other names could not be categorised with respect to their individual roles/occupations because the corpus did not contain enough examples related to specific roles/ occupations. In addition to this, these names were not expected to be present in large numbers, and therefore all these examples needed to be identified and categorised as 'Other Name'. Rules were developed via contextual analysis of Development set for the identification and anonymization of 'Other Name'.

For example,

Seen by Nurse practitioner Lara Jones. // Other Name

Split up with partner, Mark.

Although, there were very few examples of 'Other Name' found in the corpus the dictionary of 'others' along with rules will be able to provide strong clues for the identification of 'Other Name'. Example rules are shown in Figure 3-13. Similarly, there were few names of places found in the corpus but the general rules were developed on the basis of these examples, explained in the next section.

```
Rule: Other Names
(
  (NAME)
  (SPACE)
  (NAME)
  {Token.kind==punctuation}
  {Lookup.majorType == other}                    //Dictionary of roles/occupation other than doctor titles.
  )                                              //identify 'Lara Jones, Nurse'
  |
  (
  {Lookup.majorType == other}
  {Token. kind==punctuation}
  (SPACE)
  (NAME)
  (SPACE)
  (NAME)
  )                                              //identify 'husband, John Smith'
)
: label
-->
: label.OtherName= {Rule=OtherNames}
```

Figure 3-13: Example rules for identification and classification of Other Name.

3.4.2.1.4 Identification and classification of ‘Place Name’

As mentioned earlier, the corpus contained few but interesting examples of place names which appeared with general terms (general practice, hospital, group practice, etc.). For instance, 'Headingley group practice' was a place name in the corpus in which 'Headingley' is the identification of place. This leads to an observation that any other city name associated with general terms can determine another place name such as 'Meanwood group practice', 'Sherburn group practice', 'Yaxley group practice', etc. These general terms do not identify any personal

health information in medical narratives and should be kept in the records. The anonymization of place names can be considered as completed by the identification and anonymization of 'Headingley' with its PHI category. In this way, these general terms will also help to maintain the readability of the text for analysis, as shown in Figure 3-14.

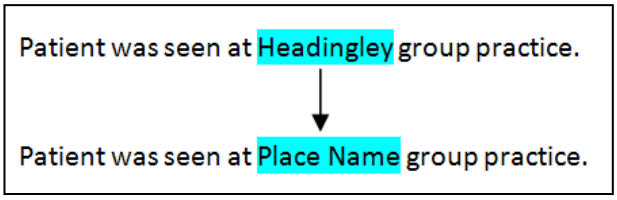


Figure 3-14: Anonymization of place name.

Therefore, the existing dictionary of place names was updated with all places excluding any general terms (hospital, university, bus station, etc.). This updated dictionary helped in the identification of single word place names by applying string matching from the dictionary. On the other hand, it was noticed that some of place names were wrongly identified by the baseline system, shown in Figure 3-15 and therefore rule-patterns were developed to restrict wrong place names.

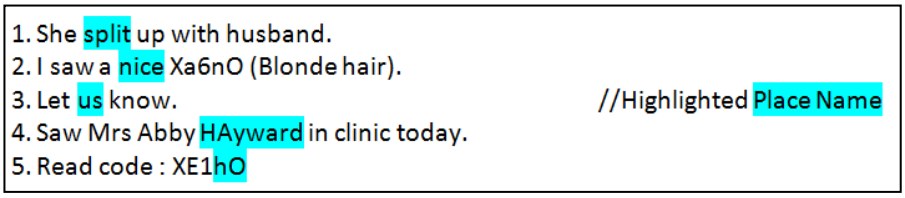


Figure 3-15: Issues identified in the identification of place names using the baseline system.

For examples 1, 2, 3 and similar cases, a rule including patterns was added to restrict the orthographic feature of 'lowercase', and for cases such as examples 4 and 5, a pattern was added to restrict the orthographic feature of 'mixedCaps' shown as follows.

```
Rule: PlaceName
(
{ lookup. majorType==place,          //Dictionary containing place names.
  Token.orth!=lowercase,
  Token.orth!=mixedCaps
```

}
):label
→
:label.PlaceName={Rule=PlaceName}

To deal with the cases that include 'READ codes' (such as example 5 in Figure 3-15), orthographic feature 'mixedCaps' was used in the pattern to restrict incorrect identification of place names.

The identification and classification of PHI categories achieved 100% of f-measure on the development corpus and performance measurements on each category are shown in Table 3-6. The formulas used for the calculation of precision, recall and f-measure are provided in Section 3.5.

Table 3-6: Performance measurements achieved on Development set.

PHI categories	**Precision (%)**	**Recall (%)**	**F-Measure (%)**
Patients Name	99	100	100
Doctors Name	100	100	100
Other Name	100	100	100
Place Name	100	100	100
Micro Summary	99	100	100

3.4.2.2 Anonymization of PHI by replacing them with their respective PHI categories

After the development of the sub-module for identification and classification of PHI categories, the next step was to anonymize names associated with their respective PHI categories. In general, these PHI categories can simply be removed to complete the process of anonymization or alternatively can be replaced by non-identifiers ('ABC', 'XXX', etc). However, in the present study, the output of identification of PHI categories was first exported in XML format using an option available in the GATE tool shown in Figure 3-16.

<paragraph>
Mrs
<PatientsName>Zara Turner</PatientsName> attended her New Patient consultation. She is a credit retrieval manager and is married to her husband, <OtherName>Riley</OtherName>. They live together. Appeared well. Spoke about alcohol consumption which appears quite high (about 50 units week). Typical drinking habit is 2-3 glasses of wine with her husband each evening and a couple of pints of beer at the weekend. Keen to try and reduce this. Consumption has increased due to stressful job and not having time to go swimming which is how she has relaxed in the past. Spoke to her about trying walking as a means of relaxing instead. Asked about smoking cessation. Was given advice and leaflets. Will review in couple of weeks to start a management plan. Seen by Nurse practitioner<OtherName>Lara Jones</OtherName>.
</paragraph>

Figure 3-16: Identification and classification of PHI categories exported in XML format.

Then, a Python program was written to replace the identified PHI with their respective PHI categories to complete the anonymization, shown in Figure 3-17. The replacement of PHI with their respective PHI categories will help researchers to understand the context of the corpus for further investigations.

Mrs patient name attended her New Patient consultation. She is a credit retrieval manager and is married to her husband, other name. They live together. Appeared well. Spoke about alcohol consumption which appears quite high (about 50 units week). Typical drinking habit is 2-3 glasses of wine with her husband each evening and a couple of pints of beer at the weekend. Keen to try and reduce this. Consumption has increased due to stressful job and not having time to go swimming which is how she has relaxed in the past. Spoke to her about trying walking as a means of relaxing instead. Asked about smoking cessation. Was given advice and leaflets. Will review in couple of weeks to start a management plan. Seen by Nurse practitioner other name.

Figure 3-17: Final output after anonymization of PHI with their respective PHI categories.

3.5 Evaluation

For the evaluation of identification and classification of PHI, standard information extraction metrics of precision, recall and f-measure were used (Sokolova and Lapalme 2009). Formulas of Precision, Recall and F-measure are given in this section but details of formulas are explained in the evaluation chapter of this thesis (Chapter 7).

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

Here;

tp = True positives: Correct PHI that should be identified.

fp = False positives: PHI that should not be identified.

fn= False negatives: PHI that should match but did not match by the application.

The evaluation was done against the human-annotated gold standard Evaluation set (85% of whole corpus) and achieved overall f-measure of 99%. The performance measurements for each individual PHI category are shown in Table 3-7 and details of true positives, false negatives, false positives and partial matches against gold standard annotations are provided in Table 3-8.

Table 3-7: Identification of PHI categories evaluated against Evaluation set.

PHI categories	**Recall (%)**	**Precision (%)**	**F-Measure (%)**
Patients Name	100	99	100
Doctors Name	100	100	100
Other Name	80	80	80
Place Name	92	92	92
Micro Summary	100	99	100

In comparison with the baseline results, the rule-based system for identification, classification and anonymization of PHI categories improved 24% in overall f-measure, as shown in Figure 3-18.

Table 3-8: Details of performance measurements for each PHI category on Evaluation set.

PHI categories	**Gold standard annotations**	**True positives (tp)**	**False negatives (fn)**	**False positives (fp)**
Patients Name	2117	2109	1	11
Doctors Name	6	6	0	0
Other Name	5	4	1	1
Place Name	25	24	2	2
Micro Summary	2153	2143	4	14

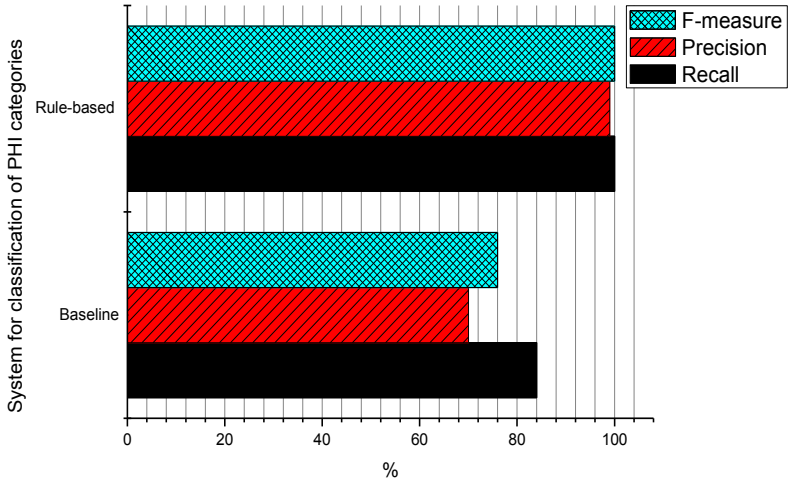


Figure 3-18: Comparison of the rule-based system with the baseline system.

3.6 Discussion

The anonymization module developed in this research scored 100% of f-measure on the Evaluation set that contained a mixture of natural language text and clinical codes. Along with this success, some false positives and false negatives were encountered during the final evaluation of the sub-module for identification and classification of PHI. These false positives were mainly because of problems that were not tackled in this study. For instance, some of the names that were not identified by the rule-based system were misspelled names. Misspellings of names were not studied in this research which is one of the limitations of this system. However, spell-checking had been studied by other researchers for other cases such as (Lew and Mitton 2012).

Similarly, in case of place names, some place names refer to local names of buildings such as 'Worsley building'. These names were missed by our system because the dictionary of place names did not contain names of local buildings and it was only able to identify place name associated with any city or country (such as Leeds General Infirmary, Bradford General Infirmary, etc.). This is one of the limitations of our system.

It was also observed that the medical students did not follow the proper format in writing consultation notes which in some cases lead to the identification of wrong place names. For

instance; in the sentence, 'Seemed Nice.', 'Nice' was wrongly identified as place name because of capitalisation. This is due to the fact that place names were mainly identified using dictionaries and rules were not developed using POS tags. This was also recognised as a limitation of this system.

Also the category ‘date’ was left out because the examples associated with ‘date’ did not provide any clue about an individual (such as patient, doctor, etc.). These dates were written in incomplete formats in the dataset, such as 'Seen in 2001', 'Scan in Jan 2001'. These formats of dates did not identify any individual, and therefore were not dealt as PHI category in this research.

Moreover, we also think that using the SnoMedTagger developed in this research to extract all medical terms is another approach of anonymizing data because this extraction will only leave terms that do not include PHI. However, this approach may not fit into research which is based on a contextual analysis of natural language datasets. For contextual analysis, the SnoMedTagger can be used to restrict medical terms which were wrongly identified as PHI.

3.7 Summary

This chapter outlined the development of a module for the anonymization of PHI. This anonymization module was used to anonymize Test dataset 1 in preparation for the evaluation of semantic tagger in this research (Chapter 8). For this anonymization module, first a corpus (Test dataset 1) was annotated with four PHI categories (Patient Names, Doctor Names, Other Name, and Place Name). This was done to develop a gold standard Development set and Evaluation set for the anonymization module. The anonymization module was completed by developing two sub-modules.

1) A rule-based sub-module for the identification and classification of four PHI categories. This sub-module was developed using 15% of the gold standard Test dataset 1 and was evaluated on 85% of gold standard Test dataset 1.

2) A Python program for the anonymization of PHI by replacing the identified PHI with their respective PHI categories.

Lastly, the results of this anonymization module were also compared with the baseline system. The anonymization module outperformed the baseline system by achieving f-measure of 100% which was 24% higher than that achieved by the baseline system.

Chapter 4. SNOMED CT® clinical healthcare terminology

4.1 Introduction

In the medical domain, Electronic Health Record (EHR) systems contain a wealth of critical information about medical concepts. This information needs to be interpreted by other healthcare professionals and therefore should be shared with consistency.

To avoid potential risks of misinterpretation and inaccuracy of clinical information, the International Health Terminology Standards Development Organisation (IHTSDO) develop and maintain Systemised Nomenclature of Medicine – Clinical Terms (SNOMED CT), clinical healthcare terminology and other clinical terminologies. In this research, SNOMED CT clinical healthcare terminology was used as a resource in the development of a semantic tagger for medical narratives.

SNOMED CT is the combination of two well-known clinical terminologies; SNOMED Reference terminology (SNOMED RT), developed by the College of American Pathologists (CAP), and Clinical Terms Version 3 (CTV3), developed by the National Health Service (NHS) in the United Kingdom (Stearns et al. 2001). SNOMED CT was selected because it is the most comprehensive multilingual clinical healthcare terminology which is widely used in the world (NLM 2011).

In addition to this fact, the National Center for Biomedical Ontology (NCBO) Ontology Recommender service was also used for the selection of the best clinical terminology for the corpus of medical narratives used in this research. This is a biomedical ontology recommender, which suggests the most appropriate ontology for annotating the relevant data (Jonquet, Musen and Shah 2010). It takes the decision on the basis of three criteria; ontologies that cover most terms/concepts present in the input text, mapping between ontologies, and size of ontologies.

A small part of the corpus was tested using NCBO Ontology Recommender service, and SNOMED CT was suggested out of more than 200 biomedical/medical terminologies for the corpus of the medical narratives used in this research. The SNOMED CT clinical healthcare terminology is distributed by the US National Library of Medicine[6] (NLM) which supports research and development in the biomedical and medical domain. The NLM also provide access to healthcare databases such as MeSH, UMLS, and MEDLINE, and among these healthcare databases SNOMED CT is distributed as part of the Unified Medical Language system – UMLS (*Unified Medical Language System® (UMLS®)*) on the basis of a valid UMLS license. The UMLS is a meta-thesaurus that contains several medical/biomedical vocabularies for their interoperability between applications. SNOMED CT is also distributed as part of and a UMLS licence was required to access SNOMED CT files. Therefore, a UMLS license was requested and granted to use SNOMED CT files used in this research.

This chapter contains the description of the SNOMED CT clinical healthcare terminology and its components, the extraction of SNOMED CT semantic categories and the use of these semantic categories in the implementation of the baseline system for this research. Finally, this chapter explains the medical semantic tag set derived from SNOMED CT to be used in this research.

4.2 SNOMED CT healthcare clinical terminology and its components

The SNOMED CT clinical healthcare terminology can be used to code or retrieve medical concepts and to analyse clinical information. It can also help to link data from different healthcare systems with standard and consistent code information. SNOMED CT is designed in the form of a hierarchy which contains top-level concept classes. These top-level concept classes are further divided into their sub-classes (Coiera 2003). Each top-level concept class and sub-class under the SNOMED CT hierarchy represents a semantic category and is implemented by three basic components.

[6] http://www.nlm.nih.gov/

- **'Concept' table:** The SNOMED CT concept table contain 386,020 concepts (SNOMED CT Version 2011). Each concept in this table has a unique name which is called the 'Fully Specified Name' (FSN) of that concept. Each FSN is associated with its semantic category written in parenthesis. For example, Entire Heart (Body Structure).
- **'Description' table:** Each unique concept in the SNOMED CT concept table has other names (synonyms, abbreviations, etc.) which are stored in the description table. For instance, 'Heart attack' is a synonym of the concept 'Myocardial Infarction' which can also be abbreviated as 'MI'.
- **'Relationship' table:** The SNOMED CT concepts are linked together by means of logical definition. The relationship table contains information to link the SNOMED CT concepts. For instance; 'Fracture of right foot' has a relationship with 'Fracture of foot'.

In this research, all the concepts were extracted from the 'Concept' table of SNOMED CT terminology (section 4.2.1). These concepts were then used as base vocabulary. The 'Description' table and 'Relationship' table were not used as base vocabulary in this research because of the limitation of the SNOMED CT terminology to identify semantic information in the medical narratives. However, the concepts in 'Description' table were used for searching the equivalent multiword concepts that were written differently in the Development dataset.

4.2.1 Extraction of SNOMED CT semantic categories

The SNOMED CT concepts were extracted from the 'Concept' table to investigate the use of SNOMED CT concepts for the identification of concepts and their classification with respective semantic categories. Initially, concept extraction was investigated on the corpus of medical narratives by using SNOMED CT concepts without their classification with semantic categories The approach for doing this is more fully described in (Hina, Atwell and Johnson 2010a). This concept extraction showed that SNOMED CT can be used to extract individual concepts. After analysing the use of SNOMED CT for concept extraction, the 'Concept table' was pre-processed for the extraction of concepts with their respective categories. The 'Concept' table, which was a tabs delimited file, contained the following attributes; CONCEPT ID, CONCEPT

STATUS, FULLY SPECIFIED NAMES, CTV3 ID, SNOMED ID, IS PRIMITIVE. The example of 'Concept' table is shown in Figure 4-1.

In the first step, a Python program was written to remove all attributes from the SNOMED CT concept file except 'FULLY SPECIFIED NAMES'. The attribute 'FULLY SPECIFIED NAMES' contained names of concepts along with their semantic categories. In the second step, another code was written to separate all the concepts with respect to their semantic categories from the attribute 'FULLY SPECIFIED NAMES' and have them to be stored in separate files. These separate files were used as dictionaries of each semantic category listing concepts. The process of separating dictionaries from the SNOMED CT concept table is shown in Figure 4-2.

CONCEPT ID	CONCEPT STATUS	FULLYSPECIFIEDNAME	CTV3 ID	SNOMED ID	IS PRIMITIVE
139784008	0	Entire tuberculum sellae (body structure)	XS10s	T-D1463	1
100419000	10	DUOVAC -M (product)	XU07K	C-D2631	1
140087001	0	Entire clivus ossis sphenoidalis (body structure)	XS1BZ	T-11183	1
100331002	10	DERMCAPS ES LIQUID (product)	XU05n	C-D2411	1
100334005	10	DERMOLAR SHAMPOO (product)	XU05q	C-D2417	1
100361005	10	DIFIL SYRUP (product)	XU06K	C-D2499	1
100362003	10	DIFIL TABS (product)	XU06L	C-D2501	1
100390004	10	DL-ALPHA TOCOPHEROL ACETATE INJECTION (product)	XU06p	C-D2569	1
10039002	0	^210m^Bismuth (substance)	XU06q	C-125B2	1
100391000	10	D-LIMONENE SHAMPOO (product)	XU06r	C-D2571	1

Figure 4-1: Example of SNOMED CT concept table.

As a result, 386,020 concepts were extracted and stored in 31 separate files (dictionaries) with respect to the 31 semantic categories (top-level concept classes and sub-classes), as shown in Table 4-1. These 31 semantic categories were then used to develop a dictionary application for the identification of concepts in medical narratives, published in (Hina, Atwell and Johnson 2010b).

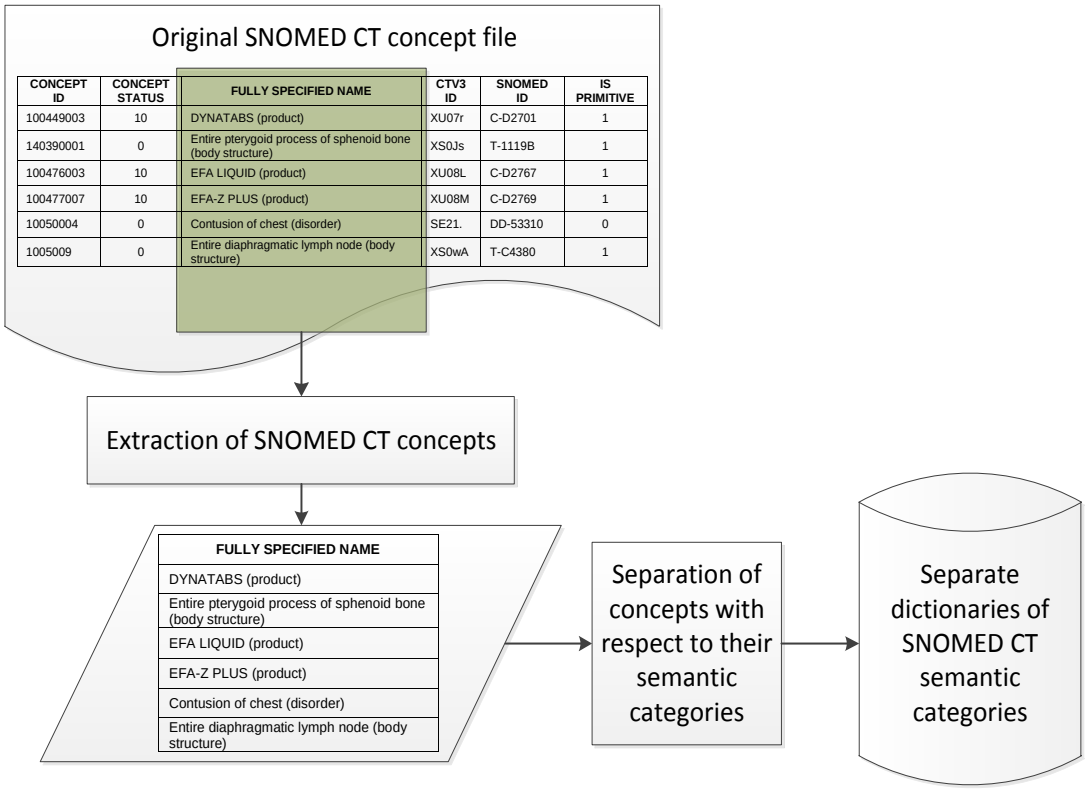


Figure 4-2: Process of extracting dictionaries of semantic categories from SNOMED CT concept table.

After the extraction of the 31 semantic categories from the SNOMED CT concept table, a simple dictionary application containing 31 dictionaries was tested on 1,176 patient records from i2b2 corpus to evaluate the frequency of concepts with respect to semantic categories. Out of 31 top-level concept classes and their sub-classes (semantic categories) from SNOMED CT, concepts associated with 15 semantic categories were not found in the corpus of medical narratives used in this research. Examples of some of these semantic categories are presented in Table 4-2. These 15 semantic categories were omitted from the research for the following reasons.

- The semantic categories such as 'Physical force', 'Religion', 'Lifestyle', 'Staging and scales', etc. were not found in the medical narratives. The concepts associated with these categories refer to special cases which are rarely used in general medical narratives.

Table 4-1: Number of concepts extracted with respect to SNOMED CT top-level concept classes and subclasses.

Semantic categories		**Number of concepts**
Top-level concept classes	**Sub-classes**	
Administrative Concept	--	75
Clinical Finding	Findings	45029
	Disorder	92492
Procedure	Procedure	73189
	Regime/Therapy	3627
Observable Entity	--	8806
Body Structure	Body Structure	26939
	Morphologic Abnormality	5127
Organism	--	35028
Substance	--	25726
Pharmaceutical/Biological Product	--	24220
Specimen	--	1359
Special Concept	Inactive Concept	8
	Namespace Concept	138
	Navigational Concept	729
Physical Object	--	5059
Physical Force	--	178
Event	--	8942
Environments/geographical locations	Environment	1250
	Geographic Location	619
Social Context	Social Concept	27
	Life style	30
	Occupation	6451
	Person	666
	Religion/Philosophy	226
Situation with explicit context	--	8538
Staging and scales	--	40
Linkage concept	Attribute	1157
	Link Assertion	8
Qualifier Value	--	10043
Record Artifact	--	294

- The concepts associated with the semantic categories such as 'Administrative concept', 'Link assertion' (For example; Has problem name, Has problem member etc.), 'Namespace concept' (For example; Extension Namespace (1000145)), 'Inactive

concept' (consists of outdated concepts, ambiguous concepts, etc.), etc. were designed to link and describe other semantic categories in SNOMED CT clinical healthcare terminology. Again, they were not considered relevant to the general medical narratives that were the motivation for this research.

Table 4-2: Examples of semantic categories that were not found in the corpus.

Missing Semantic categories	**Examples**
Staging Scales	Symptom ratings, exertion ratings, Chest pain rating...
Link Assertion	Has support, Has reason, Is etiology for,Has explanation...
Religion/Philosophy	Christadelphian movement, Jehovah's Witness religion...
Life Style	Criminal life style, Voluntary body tattooing...
Special Concept	Abnormal biochemistry finding (Navigational concept), Accidental alternative medicine overdose (navigational concept) ...

The remaining 16 semantic categories, listed in Table 4-3, were found in the corpus (medical narratives) that was used in this research. These 16 semantic categories formed the medical semantic tag set that was employed in the development and evaluation of semantic tagger for medical narratives (Chapter 6).

Table 4-3: Medical semantic tag set derived from SNOMED CT.

Tags	**SNOMED CT semantic categories**
1.	Attribute
2.	Body Structure
3.	Disorder
4.	Environment
5.	Findings
6.	Observable Entity
7.	Occupation
8.	Organism
9.	Person
10.	Physical Object
11.	Procedure
12.	Product or Substance
13.	Qualifier Value
14.	Record Artifact
15.	Regime/Therapy
16.	Situation

4.3 SNOMED CT dictionary application: Baseline system

The purpose of developing a baseline system is already described in Section 3.4.1. In order to implement the SNOMED CT dictionary application as a baseline system, 16 separate files containing concepts were used as 16 separate dictionaries of semantic categories. The baseline system used dictionaries to match exact concepts that were present in the corpus. This baseline system was set up using language processing resources in GATE software tool, as shown in Figure 4-3.

After applying basic language processing resources (Tokeniser and sentence splitter) on the corpus, the dictionaries were used to identify concepts in the corpus. The identified concepts were then classified with their respective semantic category by applying simple Java Annotation Pattern Engine (JAPE) rules (Cunningham, Mayard and Tablan 2000). JAPE rules are explained in Section 3.4.2.1.

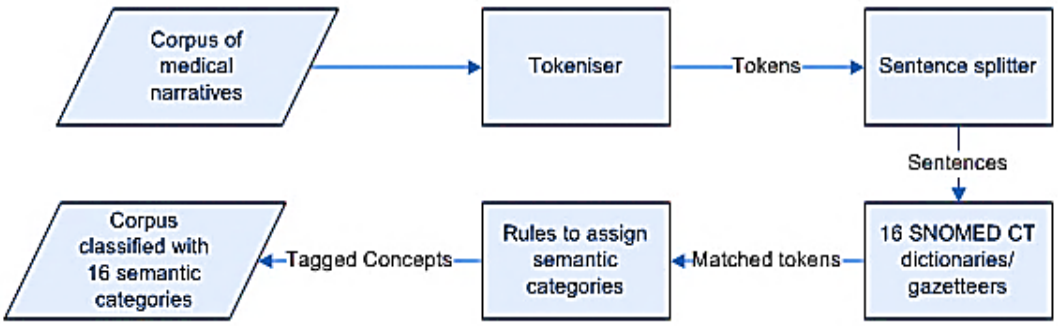


Figure 4-3: System flow of the baseline system.

For example;

Rule: BodyStructure

(

{lookup.majorType==Body}

): label

→

:label.BodyStructure= {Rule= BodyStructure}

This rule will first match concepts from the dictionary of 'Body Structure' with its majorType and then assign a label/tag of 'Body Structure' to the matched concepts. Similarly, the other 15 semantic categories were also identified and classified by the baseline system in the corpus.

Dictionaries predictably lose information; therefore for the initial investigation, Development dataset was annotated using the baseline system. The results achieved by the baseline system are presented in Table 4-4. The results indicated that the baseline system identified a very limited amount of semantic information on its own in the Development dataset. However, the semantic category 'Attribute' was an exception. This can be attributed to the fact that the semantic category 'Attribute' mostly contain single word concepts which were easily identified by the dictionary. The output produced was then manually reviewed for the identification of language issues associated with the concepts that were not identified by the SNOMED CT dictionary application (baseline system). These issues are mentioned in Table 4-5 and were also presented in (Hina, Atwell and Johnson 2011).

Table 4-4: Performance measurements of the baseline system on Development dataset.

Semantic Categories	**True positives**	**Actual concepts in Gold Standard**	**Recall (%)**	**Precision (%)**	**F-measure (%)**
Attribute	604	640	**94**	**91**	**93**
BodyStructure	75	221	34	93	50
Disorder	193	376	51	95	67
Environment	91	226	40	98	57
Findings	217	446	49	83	61
ObservableEntity	88	164	55	81	65
Occupation	34	94	36	56	44
Organism	3	7	43	75	55
Person	145	203	71	100	83
PhysicalObject	16	114	14	100	25
Procedure	199	697	29	86	43
ProductorSubstance	202	385	52	63	57
QualifierValue	886	1347	66	68	67
RecordArtifact	11	42	24	92	39
Regime/Therapy	24	102	24	89	37
Situation	22	61	36	100	53
Micro summary	2810	**5125**	55	79	65

On observing the concepts that were not identified by the baseline systems, it was concluded that the dictionaries were unable to identify complex multiword concepts, concepts which were paraphrases of the original concepts in the dictionaries, plural concepts and abbreviations of the concepts.

By considering the language issues mentioned in Table 4-5, *where semantic categories are presented in parenthesis*, the annotation guidelines were developed for non-domain users and domain experts. These annotation guidelines were used to develop the gold standard corpus of medical narratives (explained in Chapter 5).

Table 4-5: Language issues identified by SNOMED CT dictionary application.

Example concepts found in corpus which were missed by SNOMED CT dictionary application	**Equivalent concepts present in SNOMED CT vocabulary/dictionaries**
Paraphrasing Problem: Use of punctuations and linguistic features Example concept in the corpus: 'CT of the head, neck'	
1. CT (Procedure) 2. CTof the head (Procedure) 3. CT of the head, neck (Procedure)	1. CT Scan of head (Procedure) 2. CT Scan of neck (Procedure)
Abbreviation/ Acronym Problem Example concept in the corpus: 'CPAP Pressure'	
1. CPAP Pressure (Procedure)	1. CPAP treatment 2. CPAP - Continuous positive airways pressure 3. Continuous positive airways pressure therapy 4. CPAP - Continuous positive airways pressure therapy 5. Continuous positive airway pressure ventilation treatment (Regime/therapy) 6. Continuous positive airway pressure ventilation treatment (Procedure) 7. Continuous positive airway pressure ventilation treatment
Plural concepts Example concept in the corpus: 'legs'	
1. Legs (Body structure)	1. Entire lower limb 2. Hind limb 3. LL - Lower limb 4. Lower limb 5. Entire lower limb (body structure) 6. Leg
Multiword concepts (also include section headers in document) Example concept in the corpus: 'Chronic renal insufficiency'	
1. Chronic renal insufficiency (Disorder)	1. Insufficiency (Findings) 2. Chronic insufficiency (Findings)

In summary, the implementation of the baseline system (SNOMED CT dictionary application) not only provided baseline results (discussed in Chapter 7) against the human-annotated gold standard corpus (explained in Chapter 5), but were also used to analyse concepts that were not identified by the baseline system. Furthermore, this baseline system was also used to pre-annotate dictionary concepts in the gold standard corpus with the medical semantic tag set (explained in the next section).

4.4 Medical semantic tag set derived from SNOMED CT

This section contains a description of the 16 semantic categories that were derived from the SNOMED CT top-level concept classes and sub-classes. Appendix A also includes the definitions that were used by annotators and reviewers (discussed in Chapter 5).

- **Attribute**

The concepts in this semantic category represent relationships between SNOMED CT concepts. Some concepts in the 'Attribute' semantic category can be used to define concepts in a logical manner.

Example of concepts in 'Attribute': Associated with, After, Causing, Due to, During, etc.

- **Body Structure**

The concepts in this semantic category are normal/abnormal anatomical structures and also specify the body sites involved by a disease or procedure.

Example of concepts in 'Body Structure': Lung, Heart tissue, zone of lung, Polyp, etc.

- **Disorder**

The semantic category 'Disorder' is a sub-class of the top-level concept class 'Clinical Findings'. The concepts under this semantic category are diseases or disorders and always represent abnormal clinical states.

Example of concepts in 'Disorder': Tuberculosis, burn shock, bursitis of hand, Buruli ulcer, etc.

- **Environment**

The semantic category ‘Environment’ contains all types of environments and locations.

Example of concepts in ‘Environment’: Home, Emergency department, Warehouse, I.C.U., Zoo, etc.

- **Findings**

Like ‘Disorder’, ‘Findings’ is also a sub-class of the top-level concept class, ‘Clinical Findings’. The concepts under this semantic category are the results of clinical observations or examinations and include normal as well as abnormal clinical states.

Example of concepts in ‘Findings’: Able to run, Absence of toe, Anxiety, Death, etc.

- **Observable Entity**

The concepts in this semantic category represent questions or procedures which can produce an answer or a result. These entities can also be used as an element where a value can be assigned.

For instance, ‘Left ventricular end-diastolic pressure (Observable Entity)’ could be interpreted as the question, “What is the left ventricular end diastolic pressure?” or “What is the measured left ventricular end-diastolic pressure?”

Observables are entities that could be used to code elements on a checklist or any element where a value can be assigned. For instance, ‘Colour of nail’ is an observable, whilst ‘Grey nails’ is a finding.

One use for ‘Observable Entity’ in a clinical record is to code headers on a template. For example, ‘Gender (Observable Entity)’ could be used to code a section titled “Gender” where the user would answer “male” or “female”. These values of “Gender” would then constitute a finding.

Example of concepts in ‘Observable Entity’: ‘colour of nail’, ‘age’, ‘gender’, ‘length of ulna’, ‘blood pressure’, etc.

- **Occupation**

It is a sub-class of the top-level concept class ‘social context’ and contains all concepts which are occupations.

Example of concepts in ‘Occupation’: ‘doctor’, ‘general practitioner’, ‘nurse’, ‘clerk’, ‘manager’, ‘actor’, etc.

- **Organism**

The concepts in this category include organisms of significance in human and animal medicine or in modelling the causes of diseases.

Example of concepts in ‘Organism’: ‘algae’, ‘alnus’, ‘amoeba’, ‘black fly’, ‘cryptocotyle’, etc.

- **Person**

Like ‘Occupation’, it is another sub-class of the top-level concept category ‘social context’ and contains concepts which can be referred to as a person.

Example of concepts in ‘Person’: ‘employer’, ‘patient’, ‘baby’, ‘father’, etc.

- **Physical Object**

Concepts in this semantic category include natural or man-made objects or objects used to model the concepts in the semantic category ‘Procedure’.

Example of concepts in ‘Physical Object’: ‘book’, ‘needle’, ‘boiler’, ‘cloth’, etc.

- **Procedure**

The concepts in this category include activities performed in the provision of health care.

Example of concepts in ‘Procedure’: ‘radiography’, ‘measles vaccination’, ‘operation on the ear’, ‘optimal surgery’, etc.

- **Product or Substance**

For the present study, two top-level concept classes ‘pharmaceutical/biological product’ and ‘substance’ were combined to form this semantic category. This was done on the basis of the

observation by domain experts that these two semantic categories (concept classes) were interchangeably used in the medical narratives and mostly used to record ‘Medications’. However, in the original SNOMED CT hierarchy, the semantic category ‘pharmaceutical/biological product’ contained names of drug products and the semantic category ‘substance’ contained chemical constituents of drug products (in the ‘pharmaceutical/biological product’ category), food and chemical allergens and adverse reactions and toxicity information

Example of concepts in ‘Product or Substance’: ‘vancomycin’ (Product), ‘VAL syrup’, ‘topical from Zinc’ (Product), sodium citrate (Substance), etc.

- **Qualifier Value**

The semantic category ‘Qualifier Value’ contains some of the concepts used as values for SNOMED CT attributes that are not present elsewhere in SNOMED CT. Such a code may be used as the value of an attribute in a defining relationship in pre-coordinated definitions, and/or as the value of an attribute in a qualifier in a post-coordinated expression. However, the values for attributes are not limited to this hierarchy and are also found in hierarchies other than the ‘Qualifier value’.

Example of concepts in ‘Qualifier Value’: ‘left’, ‘right’, ‘first’, ‘upper’, ‘unit of rate’, ‘simple’, etc.

- **Record Artifact**

The ‘Record Artifact’ concepts are entities created by a ‘person’ to provide information on events or records.

Example of concepts in ‘Record Artifact’: ‘death summary’, ‘discharge summary’, ‘summary report’, ‘radiology report’, etc.

- **Regime / Therapy**

It is a sub-class of the top-level concept class ‘Procedure’ and includes concepts focal in the ‘Procedure’.

Example of concepts in ‘Regime Therapy’: ‘art therapy’, ‘cold therapy, ‘ear care’, dying care’, etc.

- **Situation**

The concepts in ‘Procedure’ and ‘Clinical Findings’ which are one of the following types are ‘Situation’ concepts;

• Conditions and procedures that have not yet occurred.

• Conditions and procedures that refer to someone other than the patient.

• Conditions and procedures that have occurred at some time prior to the time of the current entry in the record.

Example of concepts in ‘Situation’: ‘history of anaemia’, ‘family history’, ‘no nausea’, ‘Endoscopy arranged’, etc.

4.5 Summary

This chapter explains the purpose of using SNOMED CT healthcare terminology, description of SNOMED CT and its basic components and the method developed for extracting semantic categories from the original SNOMED CT concept table.

After the extraction of semantic categories from the concept table, a simple baseline system (SNOMED CT dictionary application) was tested to select appropriate semantic categories for the corpus of medical narratives used in this research. Out of 31 top-level concept classes and sub-classes (semantic categories) 16 semantic categories were considered appropriate for the medical narratives used in this research. These 16 semantic categories were then used as dictionaries in the baseline system for the identification of concepts and their classification with the respective semantic categories. The baseline system did not only provide baseline results but

was also used to pre-annotate the corpus for the development of the gold standard (semi-automatic approach explained in the next chapter).

Chapter 5. Corpus and gold standard datasets

5.1 Introduction

In computational linguistics, the language engineering modules/applications (tokenisation, sentence splitter, named entity recognition, etc.) require annotated data for evaluation. To annotate this data, domain expertise is required for specialised domains such as the medical domain. The definition of domain expertise can vary from one researcher to another depending on the subjective domains (medical, chemistry, education, etc.).

While a growing number of natural language processing (NLP) research projects have worked on specialised domains, it is difficult for the non-domain researchers (language researchers) to work on medical corpora without the involvement of domain experts. One of the more interesting medical corpora originates from real-time EHR systems in the form of clinical documents (discharge summaries, progress notes etc.) which contain information in the form of narratives written by clinicians using a mixture of natural language and more technical medical language. For certain research tasks, particularly where a large corpus requires computation based research, these medical narratives need to be annotated.

In natural language processing research, the term ‘annotation’ means the identification of required information with its specific type/category (Part-of-speech categories, sentences, named entities such as ‘person’, ‘place’, etc.). These types/categories vary from one research objective to another. This research is aimed to help automate the process of identification and classification of semantic information in medical narratives. Therefore, the types/categories of most interest are the semantic categories specific to medical narratives (for example; ‘Disorder’, ‘Findings’, ‘Procedure’, etc.). For effective annotation, the concepts present in the corpus of medical narratives should be identified and classified with their respective semantic categories including those written by clinicians (which involve more technical medical language). This automation for the identification and classification of semantic information in medical narratives needs to start with annotated datasets for development and evaluation.

As a non-domain user, difficulties were experienced in attempting the semantic analysis of medical narratives in this research. It was thought that the use of controlled vocabularies, such as SNOMED CT, could be employed for the analysis of domain knowledge. However, such resources also have some limitations on different datasets (Friedman et al. 2001). Therefore, there was a need for a comprehensive and generic annotation scheme that could be used on medical narratives by both domain users and non-domain users.

In this chapter, we firstly explain the selection of the corpus for the development and evaluation of the SnoMedTagger. Then, the annotation guidelines for the development of a gold standard are described and the experiments and evaluation carried out following these annotation guidelines are reported. Lastly, we conclude the evaluation by presenting the inter-annotator agreement results and the final gold standard that was used for the development and testing of the SnoMedTagger developed in this research.

5.2 Selection of development dataset and test datasets

In the medical domain, the availability of data for research is always limited because of ethical reasons and access restrictions. Only a few organisations allow access to data for research, and this is often subject to participation in a challenge (*International Challenge: Classifying Clinical Free Text Using Natural Language Processing.* ; Pestian et al. 2007; *i2b2: Informatics for Integrating Biology & the Bedside*), generally to tackle a specific research question (Uzuner, Luo and Szolovits 2007; Uzuner et al. 2008a).

In this research, datasets were obtained from two different resources. The first dataset was obtained by participating in a global natural language processing challenge i2b2 for identification and classification of concepts. For this task, we implemented simple rules using SNOMED CT dictionary concepts that overlapped noun phrases for the identification of concepts, but unfortunately did not report any scores on the classification (Hina et al. 2010). The identification of noun phrases concepts was the requirement of this challenge. Moreover, the overlapped noun phrases were not found to be useful in the classification of complete semantic information in medical narratives. The i2b2 corpus was annotated by challenge

organisers with limited semantic categories (‘Problem’, ‘Treatment’ and ‘Test’) which were not following the medical semantic tag set used in this research. By participating in this NLP challenge, the main objective was not to win the task but to obtain an appropriate dataset for the development of a semantic tagger. The data selected from i2b2 corpus for the development of a semantic tagger was named as ‘Development dataset’. This dataset was most appropriate for the development of our semantic tagging application because it was reviewed by the i2b2 challenge committee and contained de-identified clinical documents (discharge summaries, progress notes) from different healthcare providers; Beth Israel Deaconess Medical Centre (MIMIC II database) and University of Pittsburgh Medical Centre. The corpus was written in US English and different healthcare partners also provided a range of variation in formatting such as capitalisation of section headers, use of punctuation, and lexical patterns (use of paraphrases, long multiword concepts). 25 clinical documents were selected from the dataset contributed by one of the healthcare provider for the development of the SnoMedTagger and 40 clinical documents from different healthcare provider’s datasets were used for testing the SnoMedTagger.

In this thesis, the dataset containing 40 documents was named ‘Test dataset 2’ because the gold standard annotations for this Test dataset 2 were completed after ‘Test dataset 1’ (described in Chapter 3). The reasons for using two test datasets are tabulated in Table 5-1.

The second corpus ‘Test dataset 1’ (explained in Section 3.3) was written by medical students using the UK’s English. The medical students showed noticeable variation in writing up consultations, and therefore this dataset was deemed most appropriate to evaluate the general applicability of the rule-patterns of the SnoMedTagger on a different dataset. The complete corpus measurements of all these datasets are provided in Table 5-2.

It is worth noting that the corpus is comparable in size to others used in corpus annotation research, for example the Quran Annotated Corpus (Dukes and Atwell 2012) (Dukes, Atwell and Habash 2013), Swedish corpus of clinical notes (Skeppstedt, Kvist and Dalianis 2012) and the Spoken English Corpus (Brierley and Atwell 2010).

Table 5-1: Reasons for choosing test datasets from different resources.

Test datasets	Language	Methods suitable for evaluation	Reasons
Test dataset 1	English (U.K.)	Rule-based	1. Rule-based method needs to be tested on different and unseen dataset. Therefore, this Test dataset was most appropriate because it was written in English (U.K) which was different from the language analysed in the development of the semantic tagger (i.e. U.S English). 2. This dataset was extracted from an EHR- Electronic Health Record system that contains a variation of writing styles.
Test dataset 2	English (U.S.)	Rule-based, Support Vector Machine (SVM) - Machine learning	1. Machine learning method needs data of a similar nature for evaluation, and because the SVM based system was trained using 'Development dataset' from i2b2 corpus written in U.S. English, therefore 'Test dataset 2' from the same corpus ensured the American English. 2. Both datasets (Development dataset and Test dataset 2) were selected from different healthcare providers and therefore, were appropriate for the evaluation of the rule-based approach.

Table 5-2: Corpus measurements.

Annotations	Development dataset (from i2b2 corpus)	Test dataset 1 (SystmOne)	Test dataset 2 (from i2b2 corpus)	Total
Tokens	16380	8874	52041	77295
Sentences	749	582	2815	4146
Concepts	5125	2672	20853	28650

5.3 Development of gold standard corpus

For the development of a gold standard corpus for the medical/biomedical domain, general purpose semantic annotation platforms (such as Mechanical Turk or KIMO (Popov et al. 2003; Popov et al. 2004)) are not applicable. However, ontology-based web annotators can be used in

the annotation of medical/biomedical text with some limitations that are described in the following section.

5.3.1 Limitations in developing gold standard using existing systems

Being non-domain users, we were highly motivated to produce a gold standard corpus for medical narratives without having to rely on help from domain experts. For this reason, instead of hiring domain experts, the following well-known systems were explored for the identification of SNOMED CT concepts in medical narratives.

1. MetaMap (Aronson 2001) (Aronson and Lang 2010).

2. BioPortal web annotator (Noy et al. 2009) (Whetzel and Team 2013).

MetaMap provides "Semantic knowledge representation" of medical/biomedical text using ontologies and special modules based on regular expressions rules. These rules process the input text in the form of phrases and return concepts with the semantic categories using an interactive interface. MetaMap also allows the user to select the required ontology and semantic categories. One of the major drawbacks found when using MetaMap was that this system was developed using MEDLINE abstracts consisting of formal English language. The language in journal articles is quite different from the one used in medical narratives. Generally, these journal articles do not contain incomplete medical terms, non-standard abbreviations and variations or paraphrasing of multiword concepts. For this reason, a span of text taken from medical narratives was tested as shown in Figure 5-1. One example of the difference between medical abstracts and real world medical narratives is the use of the multiword concept, 'CT' which is a frequently used short form of 'CT scan', often written by clinicians in medical narratives, but not recognised by the MetaMap system.

This space is deliberately left blank due to pagination.

Input Text:

```
CT of the chest, abdomen, and pelvis.
```

Results:

```
Phrase: "CT"
>>>>> Phrase
ct
<<<<< Phrase

Phrase: "of the chest,"
>>>>> Phrase
chest
<<<<< Phrase
>>>>> Candidates
Meta Candidates (4):
   666 Chest [Body Location or Region]
   583 BREASTS (Breast) [Body Part, Organ, or Organ Component]
   583 Breast (Entire breast) [Body Part, Organ, or Organ Component]
   566 Mammary (Mammary gland) [Body Part, Organ, or Organ Component]
<<<<< Candidates
>>>>> Mappings
Meta Mapping (666):
   666 Chest [Body Location or Region]
<<<<< Mappings

Phrase: "abdomen,"
>>>>> Phrase
abdomen
<<<<< Phrase
>>>>> Candidates
Meta Candidates (3):
   666 Abdomen [Body Location or Region]
   666 Abdomen (Abdominal cavity) [Body Location or Region]
   666 Abdomen (Entire abdomen) [Body Part, Organ, or Organ Component]
<<<<< Candidates
>>>>> Mappings
Meta Mapping (666):
   666 Abdomen [Body Location or Region]
  <<<<< Mappings

  Phrase: "and"
  >>>>> Phrase
  <<<<< Phrase

  Phrase: "pelvis."
  >>>>> Phrase
  pelvis
  <<<<< Phrase
  >>>>> Candidates
  Meta Candidates (3):
     666 Pelvis [Body Location or Region]
     666 PELVIS (Pelvic cavity structure) [Body Space or Junction]
     666 Pelvis (Entire pelvis) [Body Part, Organ, or Organ Component]
  <<<<< Candidates
  >>>>> Mappings
  Meta Mapping (666):
     666 Pelvis [Body Location or Region]
  <<<<< Mappings
```

Figure 5-1: Output of interactive MetaMap.

MetaMap was able to identify individual concepts of 'abdomen' and 'pelvis' with the semantic category 'Body location', but was unable to identify the complete phrase 'CT of the chest, abdomen, and pelvis' with the semantic category 'Therapeutic procedure'. Our conclusion from our trials with MetaMap was that the system cannot be used to annotate complete semantic information in medical narratives. This problem was also reported by other researchers (Meystre and Haug 2005).

The second drawback of MetaMap highlighted by (Aronson and Lang 2010) is that this system has never been methodologically evaluated against a human-annotated gold standard. Another issue is that MetaMap does not provide direct mapping between SNOMED CT semantic categories and UMLS semantic categories. This was also confirmed through personal communication with Dr. Alan Aronson who developed Metamap. The only information available about mapping UMLS and SNOMED CT semantic categories can be accessed via the National Library of Medicine Webpage[7], which requires medical expertise to understand. Moreover, this information is insufficient for non-domain users to understand the complete mapping of UMLS semantic categories with the SNOMED CT semantic categories.

The second system examined was 'NCBO BioPortal web annotator' which gives a selection of more than 200 biomedical ontologies to annotate text (Whetzel and Team 2013). BioPortal provides an API facility to process large datasets in batch mode. We used the SNOMED CT ontology to process the same input text which was used to study MetaMap; the output of BioPortal is shown in Figure 5-2.

One limitation of the ontology-based BioPortal web annotator noticed was the limited and controlled language of ontology, which was insufficient to identify complete concepts written in medical narratives (Figure 5-2). This rendered BioPortal insufficient for the semantic tagging of medical narratives. However, to investigate the use of complete SNOMED CT ontology, the BioPortal system was used for evaluation against the semantic tagger developed in this research (explained in Chapter 8).

[7] http://www.nlm.nih.gov/research/umls/Snomed/snomed_represented.html/

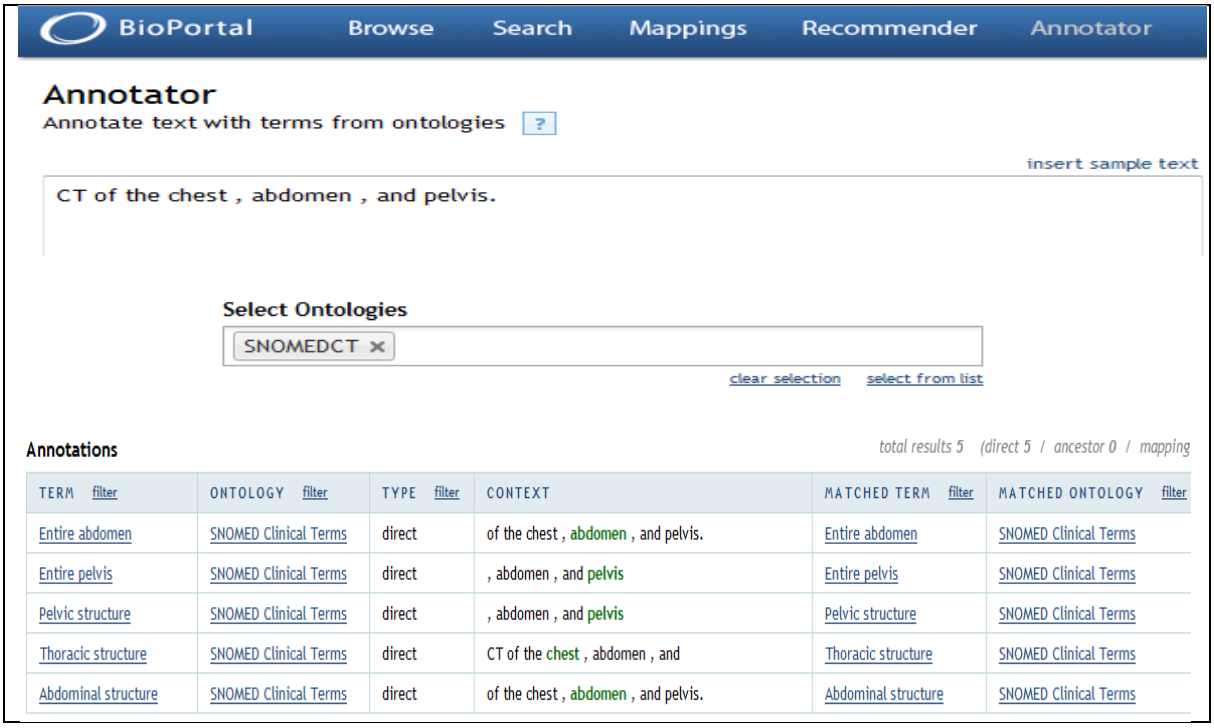


Figure 5-2: Output of BioPortal web annotator.

Other than these two well-known systems, other researchers reported the development of a gold standard dataset of 1,556 concept annotations following specific annotation guidelines to evaluate their biomedical named entity recognition system (Ogren, Savova and Chute 2008). In their research, four annotators annotated a corpus (from a Mayo Clinic repository) of 47,975 words with three annotations; 1) concepts associated with the SNOMED CT semantic category 'Disorder', 2) concept code and 3) context using annotation guidelines. They extracted 82,813 'Disorder' concepts from the SNOMED CT clinical vocabulary (*SNOMED CT User Guide, January 2011 International Release*) and used Rich Release Format - RRF browser [8] to search key words of the concepts and hierarchical navigation for the annotation of concepts. For annotation purposes, they followed two strategies; semi-automatic and manual. In the semi-automatic strategy, two annotators were provided with a corpus which was pre-annotated by the MetaMap system (Aronson 2001). This strategy was faster because annotators only had to add or remove annotations following the annotation guidelines. In the manual strategy, the other two annotators annotated the un-annotated corpus following the same annotation guidelines and

[8] http://www.nlm.nih.gov/research/umls/new_users/online_learning/UMLST_009.html

achieved overall 74.6% agreement and 82.1% agreement for overlapping spans using kappa (Cohen 1960; Carletta 1996).

In other cases, researchers working on medical/biomedical data have also spent a considerable amount of resources in designing annotation guidelines and in hiring domain experts for the annotation of required entities, see for example (Roberts A 2007; Ohta, Tateisi and Kim 2002; Wang 2007). These annotation guidelines are not generally applicable to medical narratives and were not considered appropriate for this research. Therefore, we aimed to develop general and comprehensive annotation guidelines to develop a gold standard for researchers working on medical narratives. Annotation experiments conducted in this research followed semi-automatic and manual approaches (explained in Section 5.4.2 and Section 5.4.3).

5.3.2 Annotation guidelines

Chapter 4 described the language issues identified by the baseline system (Table 4-5) which were considered to develop the annotation guidelines. These annotation guidelines were initially developed for non-domain users but were also used to guide domain experts later in this research. These annotation guidelines were based on a medical tag set derived from SNOMED CT. Other resources and the annotation tool used in the annotation of gold standard are as follows:

1. All annotations were marked using GATE- General Architecture for Text Engineering tool (Cunningham et al. 2011). GATE is an open source tool for language engineering and can be downloaded from http://gate.ac.uk/download/.
2. The SNOMED CT dictionary application (described in Section 4.3) was used for the pre-annotation of corpus with dictionary concepts associated with 16 semantic categories.
3. The Fact sheet in Appendix A was provided to annotators and reviewers for the description of 16 semantic categories.
4. SNOMED CT clinical vocabulary version 2011.

5. BioPortal web annotator developed by (Noy et al. 2009) (described above) was used to verify and annotate the remaining concepts following the annotation guidelines which are described in this section.

The manual annotation of a corpus is a time-consuming and expensive process. In comparison to manual annotation, pre-annotating the corpus automatically (using dictionaries) reduces time. Therefore, in the first step, annotators were required to load the SNOMED CT dictionary application (explained in Section 4.3) into the GATE tool. The corpus had been pre-annotated by the SNOMED CT dictionary application to help the annotator get initial annotations. This pre-annotation step is only required for semi-automatic annotation and should not be followed in case of manual approach. Each medical concept in the corpus was annotated with one or more semantic categories. However, this dictionary application was not be able identify all semantic information because the dictionaries do not contain all the concepts found in the corpus of medical narratives. This unidentified semantic information is the result of the rich expressiveness of natural language which includes paraphrases of concepts, abbreviations of concepts and complex multiword concepts often used in medical narratives. Examples of some of these were presented in Table 4-5.

After annotating the corpus from the dictionary concepts, the annotators should understand the description and examples of semantic categories given in the ‘Fact sheet’ (Appendix A). Annotators should manually read each pre-annotated document to add/remove annotations which were missed by the SNOMED CT dictionary application considering following language issues:

- Clinicians often use **complex multiword and/or overlapping concepts**; all overlapping and/or complex multiword concepts should be annotated with up to three levels of granularity as shown in Figure 5-3. The levels of granularity were decided by analysing the maximum length of multiword concepts that occurred in sentences. These levels of granularity should be followed throughout the annotation process.

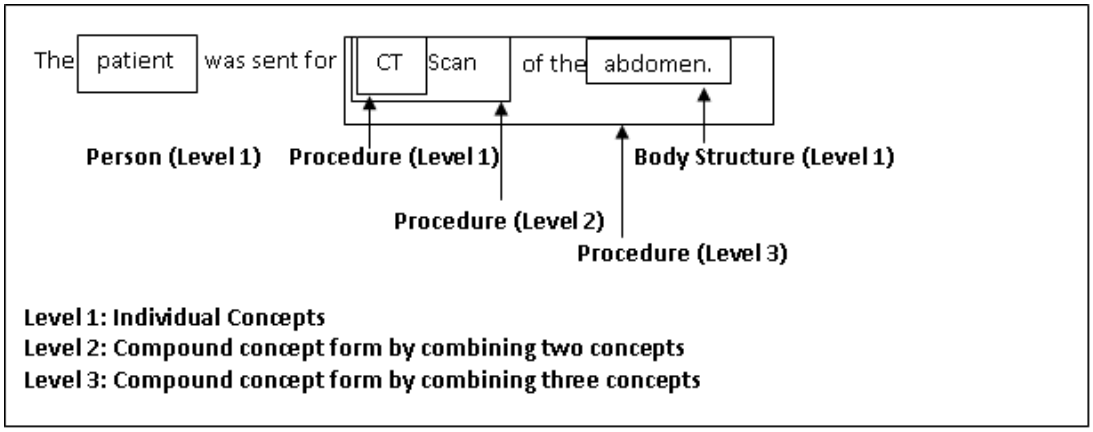


Figure 5-3: Levels of granularity to be followed in gold standard annotations.

- Clinicians also write **incomplete concepts or short names** of the concepts which will not be identified by SNOMED CT dictionary application. Table 5-3 shows examples of incomplete concepts in the corpus and their equivalent concepts in the SNOMED CT vocabulary. These concepts should be annotated by searching for keywords of the concept in the SNOMED CT clinical vocabulary.

Table 5-3: Examples of incomplete concepts and short names of concepts missed by the SNOMED CT dictionary application

Concepts missed by the SNOMED CT dictionary application	**Equivalent concepts in the SNOMED CT clinical vocabulary**
Dovonex (Semantic category= Product or Substance)	Dovonex 50 micrograms/g (Semantic category= Product or Substance)
Gestation (Semantic category= Findings)	Gestation finding (Semantic category= Findings)
Brain CT (Semantic category= Procedure)	CT of brain/ CTScan of brain (Semantic category= Procedure)

Alternatively, an annotator can also use the BioPortal web annotator to search for keywords of missed concepts as shown in Figure 5-4. On the other hand, domain experts can annotate these concepts using their own domain knowledge.

This space is deliberately left blank due to pagination.

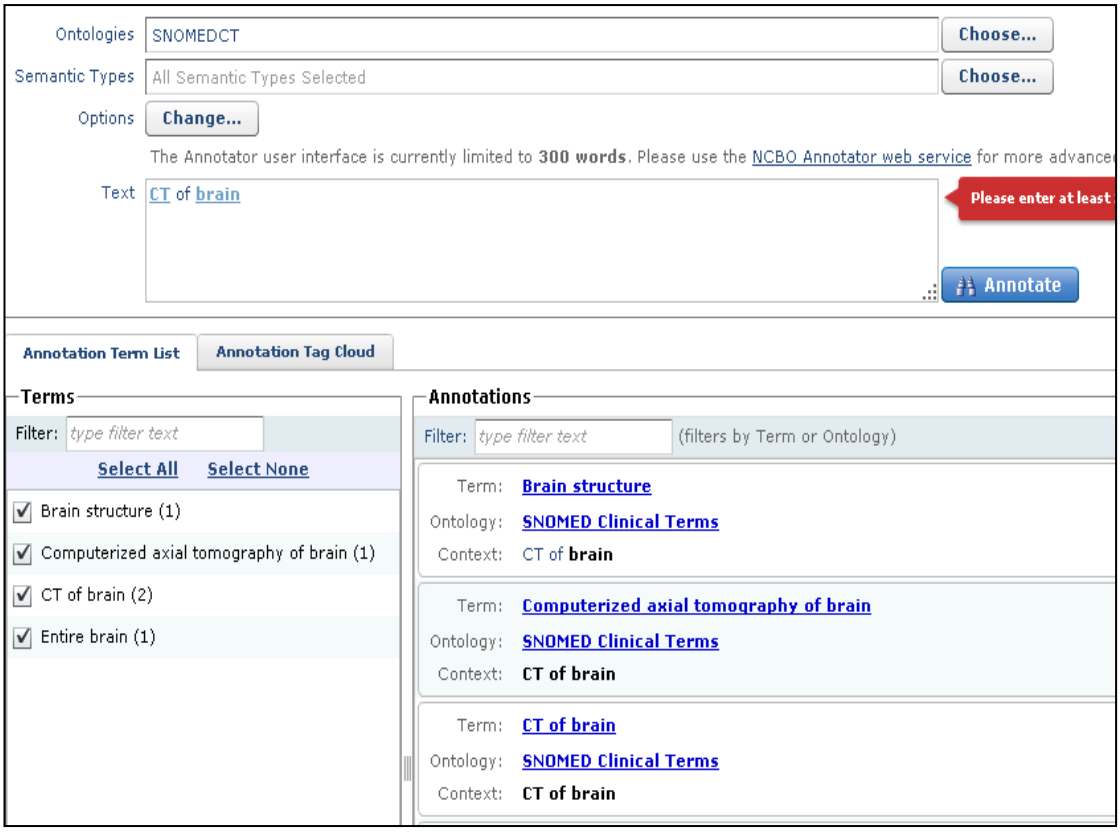
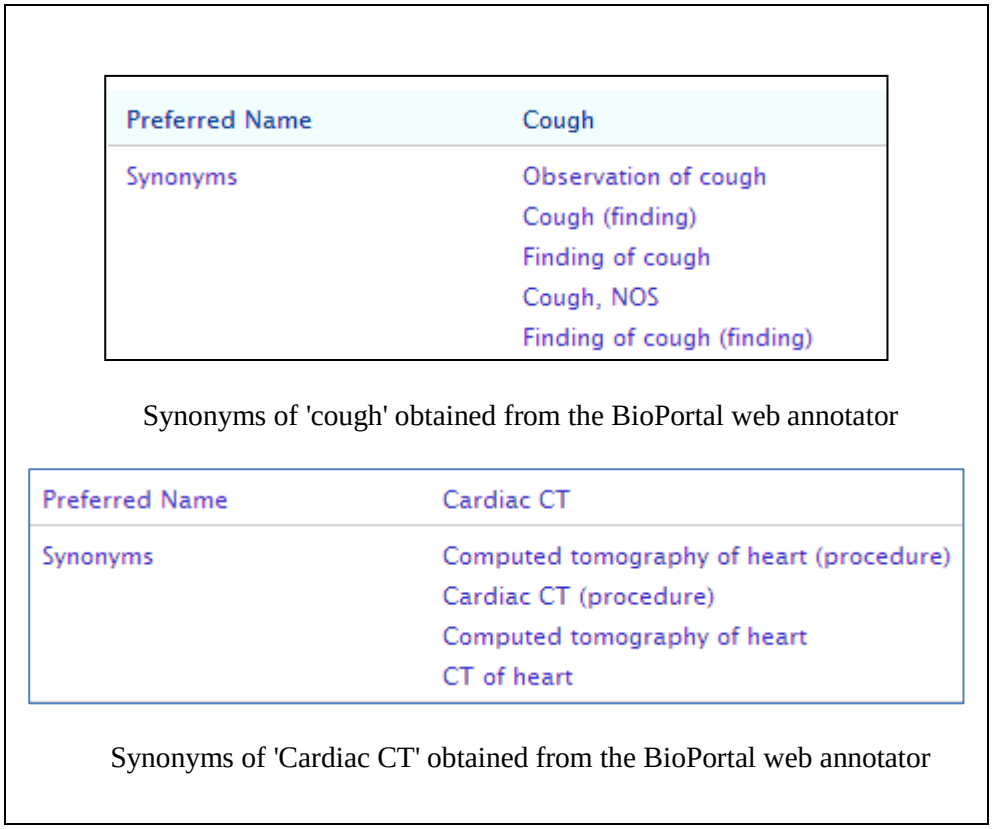


Figure 5-4: Example search results from BioPortal web annotator

- Clinicians often write **paraphrases and synonyms** of concepts in clinical documents which will not be completely annotated by the SNOMED CT dictionary application. Annotators should annotate such concepts by searching for keywords of the concepts in the SNOMED CT vocabulary or by using the BioPortal web annotator. BioPortal produces a list of synonyms on the search of a concept, as shown in Figure 5-5.
- Clinicians use **abbreviations or acronyms** as part of a **multi-word concept**. Annotators should annotate the abbreviations or acronyms of concepts individually as well as the multiword concept they appear in, as shown in Figure 5-6(a) Clinicians also write acronyms which should be annotated but are not present in clinical vocabularies. An example searched using the BioPortal web annotator is presented in Figure 5-6(b).

This space is deliberately left blank due to pagination.



Preferred Name	Cough
Synonyms	Observation of cough
	Cough (finding)
	Finding of cough
	Cough, NOS
	Finding of cough (finding)

Synonyms of 'cough' obtained from the BioPortal web annotator

Preferred Name	Cardiac CT
Synonyms	Computed tomography of heart (procedure)
	Cardiac CT (procedure)
	Computed tomography of heart
	CT of heart

Synonyms of 'Cardiac CT' obtained from the BioPortal web annotator

Figure 5-5: Examples of synonyms obtained from the BioPortal web annotator.

Considering the bulleted annotation guidelines, annotators should complete the annotation of the gold standard corpus. In case of manual approach, same annotation guidelines should be considered to produce gold standard corpus. The only difference between the two approaches (semi-automatic approach and manual approach) is that the semi-automatic approach requires pre-annotation of corpus using SNOMED CT dictionary application.

For the applicability and verification of the developed annotation guidelines, two annotation experiments were conducted to develop gold standard datasets. The experiments and evaluation are explained in the next section.

This space is deliberately left blank due to pagination.

Example concepts in corpus	Concepts which should be annotated
DVT Prophylaxis	1. DVT (Disorder)
	2. Prophylaxis (Procedure)
	3. DVT Prophylaxis (Procedure)
IV fluid	1. IV (Procedure)
	2. Fluid (Product or substance)
	3. IV Fluid (Procedure)
Head CT	1. Head (Body structure)
	2. CT (Procedure)

Figure 5-6 (a) Compound concepts containing abbreviations.

Sample Text: CT of the head, chest, and C-spine. **Where;** 'C-spine' is not an acronym in SNOMED CT vocabulary as shown by BioPortal web annotator in Figure 5.6 (b)	**Concepts which should be annotated**
	1. CT (Procedure)
	2. CT of the head (Procedure)
	3. CT of the head, chest (Procedure)
	4. CT of the head, chest, and C-spine (Procedure)
	5. head (Body structure)
	6. chest (Body structure)
	7. C-spine (Body structure)

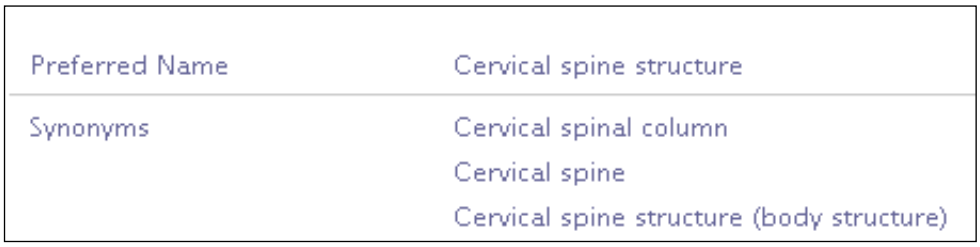


Figure 5-6 (b) Examples of synonyms of 'C-spine' searched from the BioPortal web annotator.

Figure 5-6: Examples of abbreviations or acronyms to be annotated.

5.4 Experiments and Evaluation

Before using the annotation guidelines on a large dataset it was necessary to ensure that the annotation guidelines were general and comprehensible for both domain and non-domain users. For this reason, an annotation experiment was conducted in which a non-domain user annotated the development corpus and a domain expert manually reviewed the annotations to help the validation of the annotation guidelines following the approach explained in the following

section. After the validation of the annotation guidelines, domain experts were asked to annotate the development and test datasets for this research.

5.4.1 Validation of annotation guidelines

For the validation of the annotation guidelines, the Development dataset of 16,380 tokens was annotated by a non-domain user. These annotations were then manually reviewed by a domain expert. In theory, the reliability of annotations can be measured by calculating human agreement but to save the time and cost of annotation effort, one domain expert manually reviewed the annotations and verbally agreed on more than 90% of annotations. This process of validation is shown in Figure 5-7. Note that the domain expert did not help in drafting the annotation guidelines.

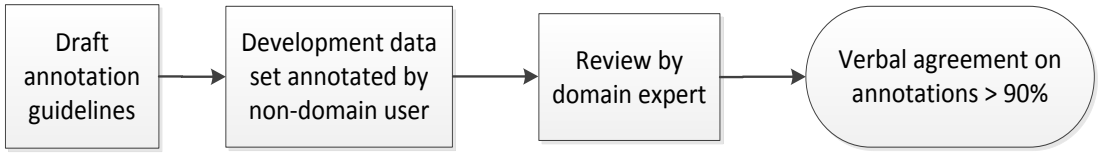


Figure 5-7 Annotation flow for validation of annotation guidelines

Two further experiments were conducted to develop gold standard datasets for this research. In these experiments, semi-automatic and manual approaches were adopted and these are explained in following sections.

5.4.2 Annotation of the Development dataset and Test dataset 1 using a semi-automatic approach

After the validation of the annotation guidelines, the Development dataset (16380 tokens) and Test dataset 1 (8874 tokens) were independently annotated by two domain experts following the semi-automatic approach described above. The semi-automatic approach was used because of limited time and resources. In this approach, the SNOMED CT dictionary application was used to pre-annotate the datasets with 16 semantic categories. The annotators were then asked to review the annotations in the corpus and add/remove annotations following the annotation guidelines (Section 5.3.2). This semi-automatic approach was found to be efficient and suitable for both domain users and non-domain users.

Inter-Annotator Agreement (IAA) is usually calculated using kappa (Cohen 1960) which was not applicable in this case (Hripcsak and Rothschild 2005). The reason is that the Kappa is calculated using $\kappa = (Pr(a) - Pr(e))/(1-Pr(e))$ where, Pr(e) is the chance agreement and Pr(a) is the observed agreement and the chance agreement needs to be calculated for distinct values which is not straight forward in this semantic annotation task (due to level of granularities).

Therefore, the inter-annotator agreement between the annotations was calculated using the Inter-Annotator Agreement (IAA), given in equation (1) which was used in annotation studies similar to ours (Roberts A 2007; Thompson et al. 2009; Kilicoglu et al. 2011). Table 5-4 shows IAA measurements for the Development dataset and Test dataset 1.

$$Inter - Annotator\ Agreement\ (IAA) = \frac{matches}{(matches + non\ matches)} \qquad (1)$$

Table 5-4: Inter-annotator agreement for Development dataset and Test dataset 1.

Annotation type/Semantic categories	**Development dataset**		**Test dataset 1**	
	Base totals	**IAA (%)**	**Base totals**	**IAA (%)**
Attribute	640	76	299	99.8
BodyStructure	221	89	25	100
Disorder	376	85	167	97
Environment	226	90	74	83
Findings	446	81	244	94
ObservableEntity	164	83	171	85
Occupation	94	96	68	100
Organism	7	100	Not present	
Person	203	99	200	100
PhysicalObject	114	90	40	98
Procedure	697	87	125	91
ProductorSubstance	385	88	222	99
QualifierValue	1347	86	848	94
RecordArtifact	42	88	17	100
Regime/Therapy	102	97	96	62
Situation	61	93	76	93
Overall	**5125**	86	**2672**	95.25

The semi-automatic approach provided high agreement scores but the disagreement was mostly due to the biased decision of annotators on the pre-annotated corpus (discussed in Section 5.4.4). Therefore, for Test dataset 2, the annotation guidelines were the subject of an experiment using a manual approach. This is explained in the next section.

5.4.3 Annotation of Test dataset 2 using manual approach

A manual approach to annotation is time consuming but in the case of the medical domain it provides more accurate annotations marked mostly using domain knowledge. This approach is not feasible in the case of non-domain users because non-domain users may easily miss many important medical and technical concepts without having any prior annotation hints.

For the annotation of Test dataset 2 (52041 tokens), two domain experts were asked to manually annotate a corpus from scratch without using the SNOMED CT dictionary application for pre-annotation. This was done to check the applicability of the annotation guidelines using domain expertise on a different dataset (Test dataset 2), and to avoid a biased decision on pre-annotated dataset.

Table 5-5: Inter-annotator agreement (IAA) for Test dataset 2.

Annotation type/ Semantic categories	**Base totals**	**IAA of Test dataset 2 (%)**
Attribute	1500	96
BodyStructure	1445	98
Disorder	1442	95
Environment	532	97
Findings	2958	86
ObservableEntity	1195	88
Occupation	222	94
Organism	57	93
Person	654	99
PhysicalObject	630	95
Procedure	2300	95
ProductorSubstance	1619	97
QualifierValue	5440	97
RecordArtifact	130	90
Regime/Therapy	408	63
Situation	321	85
Overall	**20853**	94

The inter-annotator agreement was calculated using the same formula used in the semi-automatic approach. Although, the size of Test dataset 2 was larger than the Development dataset and Test dataset 1, high inter-annotator agreement scores were achieved for each semantic category, as shown in Table 5-5.

5.4.4 An investigation of the disagreed annotations

For the disagreed annotations, we adopted the strategy of (Snyder and Palmer 2004; Girju, Badulescu and Moldovan 2006); (Roberts A 2007) suggestion and had a third domain expert to review the disagreed annotations. Table 5-6 shows the number of the disagreed annotations for each semantic category in the Development dataset and Test dataset 1 using the semi-automatic approach and Test dataset 2 using the manual approach. These different approaches were not adopted for the comparison on the same datasets, but to check the applicability of annotation guidelines on different datasets considering the availability of resources (cost and time).

Table 5-6: Total count of disagreed concepts for each semantic category in all datasets.

SNOMED CT Semantic categories (Annotation types)	**Number of disagreed concepts in each semantic category**		
	Development dataset	**Test dataset 1**	**Test dataset 2**
Attribute	31	3	99
BodyStructure	35	1	63
Disorder	70	10	125
Environment	13	2	27
Findings	98	22	663
ObservableEntity	29	4	115
Occupation	7	0	26
Organism	0	Not present	9
Person	3	0	11
PhysicalObject	12	1	54
Procedure	72	10	221
ProductorSubstance	34	4	75
QualifierValue	166	19	319
RecordArtifact	5	0	25
Regime/Therapy	6	5	201
Situation	5	7	144

For the analysis of the disagreed annotations, it was investigated that why would two domain experts assign different semantic categories to the same concept? At least four major reasons were noted as being responsible for the disagreements.

Firstly, in case of the semi-automatic approach, pre-annotation had an effect on the annotators but it was not always in the same direction. For some cases, an individual's perspective or bias played a role when pre-annotation was already done by the SNOMED CT dictionary application, while in other cases pre-annotation was found unhelpful in the assignment of semantic categories. This was due to the fact that an annotator agreed on the pre-annotated concept for some documents while assigned different semantic category for the same concept in other documents.

Secondly, the disagreement was also noticed on the basis of the semantic categories which were closely related in terms of definition in both cases (semi-automatic and manual approaches). For instance, the semantic categories, 'Findings' and 'Disorders' are subclasses of the top-level semantic class 'Clinical Finding'. Similarly, semantic categories, 'Procedure' and 'Regime//Therapy' had a very thin line of explanation between them but are separate top-level concepts in SNOMED CT and must be marked separately, which confused the annotators.

In some cases, one of the annotators also confused values of 'Findings' with 'Observable Entity' although the difference between these two categories was clearly identified in the fact sheet (Appendix A) given to the annotators. For instance, in the concept 'WBC-7.7', 'WBC' is an observable and 'WBC-7.7' is finding, but one of the annotators assigned both semantic categories (Findings, Observable Entity) to 'WBC-7.7'.

Thirdly, in both semi-automatic and manual approaches, it was noticed that some disagreed annotations were marked because of loss of concentration by the annotators after marking a number of documents. For instance, Annotator A marked 'left effusion' with the semantic category 'Findings' in most cases but did not mark the same concept in later documents. This was observed by the third domain expert during their review of the disagreed annotations.

Lastly, in the semi-automatic approach, the disagreed annotations were found because, in some cases, an annotator noticed an important concept and assigned a relevant SNOMED CT semantic category even though the concept was not present in the SNOMED CT vocabulary because the annotator felt it was important to be annotated for research purposes. For instance,

one annotator annotated the concept 'leaflet' with the semantic category 'Record Artifact' even though the concept ‘leaflet’ was not found in the SNOMED CT clinical vocabulary.

Finally, we concede that both semi-automatic and manual approaches have their own limitations, but the overall agreement score achieved by both approaches was high. Moreover, the applicability of the annotation guidelines was validated and evaluated for both types of users (domain experts and non-domain users) and the reliability of our current annotation guidelines was tested on datasets from two different resources.

5.5 The gold standard datasets

After the achievement of high inter-annotator agreement, the gold standard datasets need to be compiled. The reason for building a reliable gold standard annotated by the domain experts was to create training/development and/or test datasets to automate the process of identification and classification of semantic annotation in medical narratives (explained in Chapter 6). There are a number of principles needed to compile a dataset into a gold standard for benchmarking (Klebanov and Beigman 2009). For instance, annotators can discuss disagreed annotations (Litman, Hirschberg and Swerts 2006); in the case of more than two annotators, a majority vote strategy can decide final labels for disagreed annotations (Vieira and Poesio 2000). Or, if none of these techniques are possible then disagreed annotations can be removed from the benchmarking gold standard dataset (Markert and Nissim 2002).

As discussed in Section 5.4.4, a third domain expert reviewed and finalised the semantic category of each of the disagreed annotations for this research. Gold standard datasets were hereby compiled by constructing a consensus set from both annotation sets and by adding the disagreed annotations reviewed by third domain expert. All gold standard datasets (Development dataset, Test dataset 1, Test dataset 2) were compiled and kept separate from each other because we wanted to automate the process of semantic tagging for medical narratives using our rule-based approach.

The Development dataset was required for the development of the semantic tagger and the test datasets were required for evaluation. Table 5-7 shows the total number of SNOMED CT concepts included in the development and test datasets. To the best of our knowledge, this is the first general annotation scheme for medical narratives based on semantic categories derived from SNOMED CT.

Table 5-7: Total number of SNOMED CT concepts annotated in final gold standard.

Annotation type	**Number of SNOMED CT concepts**		
	Development set	**Test dataset 1**	**Test dataset 2**
Attribute	640	299	1500
BodyStructure	221	25	1445
Disorder	376	167	1442
Environment	226	74	532
Findings	446	244	2958
ObservableEntity	164	171	1195
Occupation	94	68	222
Organism	7	Not present	57
Person	203	200	654
PhysicalObject	114	40	630
Procedure	697	125	2300
ProductorSubstance	385	222	1619
QualifierValue	1347	848	5440
RecordArtifact	42	17	130
Regime/Therapy	102	96	408
Situation	61	76	321
Total	**5125**	**2672**	**20853**

5.6 Summary

This chapter presented the selection of datasets for the annotation of medical narratives, addressed the limitations of existing applications for annotation of gold standard corpus (medical narratives) and highlighted the issue of analysing semantic information in medical narratives. In addition to this, a comprehensive annotation scheme was described for both types of users (domain experts and non-domain users) which achieved high inter-annotator agreements (86%-95%) on datasets selected from two different resources. Moreover, the annotation scheme is based on the established medical tag set of semantic categories derived

from SNOMED CT healthcare clinical terminology and can be used on a range of medical narrative datasets. In this research, the developed gold standard datasets were used to develop and evaluate a novel automatic medical semantic tagger for medical narratives (described in Chapter 6 and Chapter 7).

Chapter 6. Semantic tagging of medical narratives using SNOMED CT

On the basis of a review of literature that is presented in Chapter 2, we identified the need for a generic and comprehensive semantic tagger which can be used for the extraction of semantic information from medical narratives. The development of such a semantic tagger, which was named as 'SnoMedTagger', formed the major contribution of this research study. In this chapter, the first section includes details of the tool(s) and the resource(s) that were required for the development of the SnoMedTagger. In the next section, the actual development process is described in detail. Lastly, the results of evaluation of the SnoMedTagger are presented.

6.1 Software tool and resources

In the development of SnoMedTagger, General Architecture for Text Engineering (GATE) was used as the software tool. Other resources that were used include the annotated Development dataset and the refined SNOMED CT dictionaries. This is explained in followings sections.

6.1.1 GATE software tool

GATE is an open source software tool which provides an infrastructure to develop and deploy software modules for language engineering (Cunningham et al. 2011). From the various GATE products, we selected GATE Developer to develop rule-based SnoMedTagger. GATE Developer can be used to construct applications for language engineering. It is a Java based graphical user interface framework that consists of language resources, processing resources and a visual resource. The language resources allow user to add or create corpus. The processing resources provide language engineering modules including tokeniser, sentence splitter, part-of-speech tagger, named entity recogniser, co-reference resolution, etc., for development and deployment of an application (Brill 1992; Brill 1994; Cunningham et al. 2002). The visual resource provides a user friendly graphical user interface to display the processing and output of a developed application.

The CREOLE - Collection of Reusable Objects in the GATE tool provide plugins for language engineering. These plugins contain a number of processing resources. In this research, the CREOLE plugins were used to carry out basic language processing tasks (such as tokenisation, sentence splitting, part-of-speech tagging), morphological analysis and to create gazetteers/dictionaries.

Moreover, Java Annotation Pattern Engine (JAPE) transducers, available in the GATE tool, were used to write rule-patterns for matching languages (Cunningham, Mayard and Tablan 2000). The JAPE language, which is based on Common Pattern Specification Language (CPSL) (Appelt and Onyshkevych 1998), and it was used to develop rule-patterns for SnoMedTagger. This is described in the following section.

6.1.2 Refined SNOMED CT dictionaries

Healthcare clinical terminologies such as SNOMED CT are hierarchical and compositional and are built from a simple set of terms/concepts that have a specific meaning. For instance, ‘diabetes’ represents a specific concept in SNOMED CT. Each concept is linked with its synonyms and with other concepts (if there is a relationship). As a result of this, the organisation of concepts is enormous and complex and it is difficult for users to find individual terms/concepts.

Like other hierarchical and compositional healthcare terminologies, SNOMED CT terminology also has the limitation of carrying redundant concepts (Sable, Nash and Wang 2001). Therefore, instead of using the complete SNOMED CT terminology hierarchy, only the ‘Concept’ table was extracted, as discussed in Chapter 4. The ‘Concept’ table comprises of unique, fully specified names of concepts. However, it also contains long multiword concepts. These long multiword concepts are comprised of individual concepts which could cause repetition in the dictionaries that were extracted from the concept file. In addition, the long multiword concepts could not be matched with the concepts written in medical narratives (due to the language issues mentioned in (Table 4-5). To avoid the above mentioned problems, long multiword concepts

needed simplification. This simplification process, which is described in the following section, is referred to as ‘refinement of dictionaries’.

6.1.2.1 Significance of refinement

While writing a medical narrative, clinicians do not follow a standard writing style. For example, one clinician may want to emphasise the severity of the problem by writing a medical description in detail whilst another may write the similar problem in a single word. For example, two ways of explaining the same problem are shown in Figure 6-1.

Clinician-1:	Clinician-2:
She has severe pain in base of left lung.	She has pain in her lungs.

Figure 6-1 Example presenting variations in the concept written by different clinicians.

In the given example, it is not necessary that the dictionaries containing long multiword concepts will identify the concept ‘pain in base of left lung’ or other paraphrases of this concept. One approach that was considered was to store every possible combination of concept phrases in dictionaries, but this was discounted on the basis that it was not practical because it cannot ensure the general applicability of dictionaries in case of medical narratives.

From our experience working with medical narratives it was quickly apparent that clinicians frequently have a preference for writing abbreviations - typically they were found to either use abbreviations or definitions, but not both, in medical narratives. This variation in writing styles causes complexity when using computing techniques to aid the identification of concepts using dictionaries.

In addition to the problem of variation in writing styles and abbreviations, the concepts in the SNOMED CT dictionaries themselves also contain some descriptions with them which do not need to be stored in the dictionaries because their description can result in inability to identify information in medical narratives. This can be understood by considering the example provided in Figure 6-1. For instance; the dictionary of the ‘Body Structure’ contains the concept ‘Lung (structure)’. This dictionary concept cannot be identified in the example. This is because of the

description '(structure)' that was associated with this concept. Therefore, the dictionaries needed refinement to resolve problems similar to the one discussed above. In the process of refinement, the dictionaries of 16 semantic categories that were derived from SNOMED CT were refined in order to develop strong base vocabulary for the SnoMedTagger.

6.1.2.2 Example cases of refinement

Chapter 4 established that the output of SNOMED CT dictionary application (baseline system) was on its own not sufficient for the identification and classification of semantic information in medical narratives. This is due to the language issues described in Table 4-5.

For instance, the concept 'CT scan of abdomen and pelvis' in medical narratives should be classified with the semantic category 'Procedure', but this concept is not present in the dictionary of 'Procedure'. Another fact is that every clinician can express this concept with any variation such as 'CT Scan of abdomen & pelvis', 'CT scan of abdomen, CT scan of pelvis', 'CT-Scan of abdomen and pelvis', 'CT of abdomen', etc. Each of these possible paraphrases for each concept cannot be stored in the dictionaries because it will only increase size of dictionaries without necessarily leading to better outcomes. To identify such concepts, generic rule-patterns are required which need strong dictionaries. For this purpose, all SNOMED CT concepts in the dictionaries that were equivalent to the concepts present in the Development dataset were refined. In addition, other multiword concepts in that contain three individual concepts (levels of granularity used in this research) were also refined. The dictionaries of semantic categories that were derived from SNOMED CT were refined in order to develop generic rule-patterns for the SnoMedTagger. Some of the more important cases of dictionary refinement are presented below. In the associated example(s), the semantic category is italicised while '→' represents the refinement process. The refinement was carried out automatically for the concepts that represent general patterns in the dictionaries (such as Case-1) while was done manually for the separation of multiword concepts (such as Case-2). For Case-3 and Case-4, Step-1 was done automatically and Step-2 was done manually. These refinement cases were also published in (Hina, Atwell and Johnson 2013b).

Case-1: Removing unnecessary words and descriptions from dictionaries.

Concepts contained information that was not written by clinicians. Such information was not required for this research. Therefore, concepts were refined by removing the unnecessary information. Examples of such concepts are as follows.

1. Removal of '[SO]' and 'NEC' from concepts where [SO] = site of origin and NEC = Not elsewhere classified

Examples;

Concept: '[SO] leg NEC – *Body Structure*'

[SO] leg NEC – *Body Structure* →leg – *Body Structure*

Concept: '[SO] thumb NEC– *Body Structure*'

[SO] thumb NEC– *Body Structure* →thumb – *Body Structure*

2. Removal of NOS from concepts where NOS= Not otherwise specified.

Examples;

Concept: 'Skin NOS – *Body Structure*'

Skin NOS – *Body Structure* → Skin – *Body Structure*

Concept: 'Nervous system NOS – *Body Structure*'

Nervous system NOS – *Body Structure* → Nervous system – *Body Structure*

3. Removal of descriptions such as '(combined site)', '(organ component)', '(structure)', 'device', etc.

Examples;

Concept: 'Joint between bodies of T7 & T8 (combined site) – *Body Structure*'

Joint between bodies of T7 & T8 (combined site) – *Body Structure* → Joint between bodies of T7 & T8 – *Body Structure*

Concept: 'Structure of mucous gland (organ component) – *Body Structure*'

Structure of mucous gland (organ component) – *Body Structure* → Structure of mucous gland – *Body Structure*

Concept: 'Vitreous membrane (structure) – *Body Structure*'

Vitreous membrane (structure) – *Body Structure* → Vitreous membrane – *Body Structure*

Concept: 'Dagger – device – *Physical Object*'

Dagger – device – *Physical Object* → Dagger – *Physical Object*

Case-2: Refinement of multiword concepts.

All the multiword concepts were simplified into individual concepts.

Examples;

Concept: 'Breast and Axillary tissue – *Body Structure*'

Breast and Axillary tissue – *Body Structure* → 1. Breast – Body Structure

2. Axillary tissue – Body Structure

Concept: 'Burn erythema of chin – *Disorder*'

Burn erythema of chin – *Disorder* → 1. Burn erythema – *Disorder*

2. Chin – *Body Structure*

Concept: 'Avulsion of nerve of eyelid – *Disorder*'

Refinement of such multiword concepts was carried out in multiple steps, as follows.

Step-1: Avulsion of nerve of eyelid – *Disorder* → 1. Avulsion – *Disorder*

2. Nerve of eyelid – *Body Structure*

Step-2: Nerve of eyelid – *Body Structure* → 1. Nerve – *Body Structure*

2. Eyelid – *Body Structure*

Since some of these individual concepts were already present in the dictionary, the duplicate concepts were removed. The refined concepts after the removal of duplicates are;

1. Avulsion – *Disorder*
2. Nerve – *Body Structure*
3. Eyelid – *Body Structure*

Concept: 'Repair of hernia of fascia of hand – *Procedure*'

Step-1: Repair of hernia of fascia of hand – *Procedure* →1. Repair of hernia- *Procedure*

2. Fascia of hand- *Body Structure*

Step-2:

a) Repair of hernia- *Procedure* → 1. Repair - *Procedure*

2. Hernia - *Disorder*

b) Fascia of hand – *Body Structure* → 1. Fascia – *Body Structure*

2. Hand – *Body Structure*

Refined concepts after removal of duplicates are;

1. Repair – *Procedure*
2. Hernia – *Disorder*
3. Fascia – *Body Structure*
4. Hand – *Body Structure*

Concept: 'Incision and exploration of rumen of stomach – *Procedure*'

Step-1: Incision and exploration of rumen of stomach – *Procedure* → 1. Incision and exploration – *Procedure*

2. Rumen of stomach - *Body Structure*

Step-2:

a) Incision and exploration – *Procedure* → 1. Incision - *Procedure*

2. Exploration - *Procedure*

b) Rumen of stomach - *Body Structure* → 1. Rumen – *Body Structure*

2. Stomach – *Body Structure*

Refined concepts after removal of duplicates are;

1. Incision – *Procedure*
2. Exploration – *Procedure*
3. Rumen – *Body Structure*
4. Stomach – *Body Structure*

Case-3: Refinement of multiword concepts containing '&/or'.

Examples;

Concept: 'Urinary tract &/or male genital organs – *Body Structure*'

Urinary tract &/or male genital organs – *Body Structure* → 1. Urinary tract – *Body Structure*

2. Male genital organs – *Body Structure*

Concept: 'Mouth &/or facial operations &/or palate operations – *Procedure*'

Step-1: Mouth &/or facial operations &/or palate operations – *Procedure* → 1. Mouth – *Body Structure*

2. Facial operations – *Procedure*

3. Palate operations – *Procedure*

Step-2:

a) Facial operations – *Procedure* → 1. Facial – *Qualifier Value*

2. Operations – *Procedure*

b) Palate operations – *Procedure* → 1. Palate – *Body Structure*

2. Operations – *Procedure*

Refined concepts after removal of duplicates are;

1. Mouth – *Body Structure*
2. Facial – *Qualifier Value*
3. Palate – *Body Structure*
4. Operations – *Procedure*

Case-4: Refinement of multiword concepts containing multiple parenthesis '[( )]'and 'or'.

Examples;

Concept: 'Hand bone: [other] or [metacarpals &/or phalanges] – *Body Structure*'

Step-1:

Hand bone: [other] or [metacarpals &/or phalanges] – *Body Structure* → 1. Hand bone - *Body Structure*

2. Other – *Attribute*

3. Metacarpals &/or phalanges – *Body Structure*

Step-2:

Hand bone – *Body Structure* → 1. Hand – *Body Structure*

2. Bone – *Body Structure*

Metacarpals &/or phalanges – *Body Structure* →1. Metacarpals – *Body Structure*

2. Phalanges – *Body Structure*

Refined concepts after removal of duplicates are;

1. Hand – *Body Structure*
2. Bone – *Body Structure*
3. Metacarpals – *Body Structure*
4. Phalanges – *Body Structure*

Concept: 'EUA Pharynx (&[oropharynx]) or [post nasal space] – *Body Structure*'

Step-1:

EUA Pharynx (&[oropharynx]) or [post nasal space] – *Body Structure* → 1. EUA Pharynx – *Procedure*

2. &[oropharynx]) or [post nasal space] – *Body Structure*

Step-2:

EUA Pharynx – *Procedure* → 1. EUA – *Procedure*

2. Pharynx – *Body Structure*

&[oropharynx]) or [post nasal space] – *Body Structure* → 1. Oropharynx – *Body Structure*

2. Post nasal space – *Body Structure*

Case-5: Adding concepts to the dictionaries.

During the development of SnoMedTagger, some concepts that were not present in the dictionaries were identified. The missing concepts were considered to be important for the identification of semantic information. Therefore, they were added to the appropriate dictionaries.

Typically such concepts were non-medical domain nouns and noun phrases. For instance, 'bus station', 'bank branch', 'ICU', 'pub', 'toilet', etc., which were added to the dictionary of the semantic category '*Environment*'. Concepts such as 'Material', 'product', 'tyre', 'pipe', 'cigarette', 'cigar', etc., were added to the dictionary of the semantic category '*Physical Object*'.

Case-6: Separation of abbreviations from their descriptions in the concepts.

The literature review identified several studies that reported the extraction of acronyms and abbreviations in biomedical text (mainly in the MEDLINE abstracts) using pattern-based approaches and regular expressions (Pustejovsky et al. 2001b; Pustejovsky et al. 2001a; Schwartz and Hearst 2003). (Nadeau and Turney 2005) adopted a supervised machine learning approach for the identification of an acronym-definition pair in a biomedical text. (Ao and Takagi 2005) presented a corpus-based algorithm for the identification of abbreviations from MEDLINE abstracts. In the present study, it was observed that clinicians generally prefer to write in either short forms (abbreviations) or long forms (definitions) of concepts while writing medical narratives. Abbreviations with their definitions are present as synonyms in the 'Description' table of the SNOMED CT ontology, but did not contain all the possible forms that could be found in the medical narratives. To deal with this problem, abbreviations and their definitions were stored separately for each respective dictionary. For instance, consider the SNOMED CT concept DVT - Deep venous thrombosis which can be written in several forms such as DVT - (Deep venous thrombosis), DVT (Deep venous thrombosis), (Deep venous thrombosis), DVT, Deep venous thrombosis, (Deep venous thrombosis) DVT, DVT (Deep venous thrombosis), (DVT), DVT: Deep venous thrombosis, Deep venous thrombosis: DVT. This concept was simplified as follows.

Concept: 'DVT - Deep venous thrombosis – *Disorder*'

DVT - Deep venous thrombosis - *Disorder* → 1) DVT - *Disorder*

2) Deep venous thrombosis -*Disorder*

This was followed by removing the duplicate concepts.

Another example is;

Concept: 'ICU – Intensive care unit – *Environment*'

ICU – Intensive care unit – *Environment* → 1) ICU – *Environment*

2) Intensive care unit – *Environment*

Abbreviation-definition pairs were not present in the Development dataset. However, a number of rule-patterns were developed to identify and classify such concepts that may be present in other datasets.

6.1.2.3 Summary of the refinement process

The refinement of SNOMED CT dictionaries was an intermediate stage in the development of the SnoMedTagger and it was conducted to construct a strong base vocabulary for generic rule-patterns. The refinement process did not significantly affect the size of the dictionaries, as shown in Table 6-1.

Table 6-1: Number of concepts in dictionaries before and after refinement process.

Dictionaries of Semantic categories	**Number of concepts in dictionaries**	
	Before refinement	**After refinement**
Attribute	1158	1170
Body Structure	26960	24833
Disorder	92496	88857
Environment	1253	1254
Findings	45039	44805
Observable Entity	8811	8752
Occupation	6451	3378
Organism	35028	35263
Person	667	489
Physical Object	5063	5117
Procedure	73201	63883
Product or Substance	49961	50023
Qualifier Value	10043	10008
Record Artifact	294	287
Regime/Therapy	3627	3048
Situation	8540	5504
Total	**368592**	**346671**

6.2 Experimental setup

This section explains the experimental setup of the rule-based SnoMedTagger that was developed to identify and classify paraphrases of concepts, abbreviations of concepts and complex multiword concepts in medical narratives. Figure 6-2 depicts the complete development process.

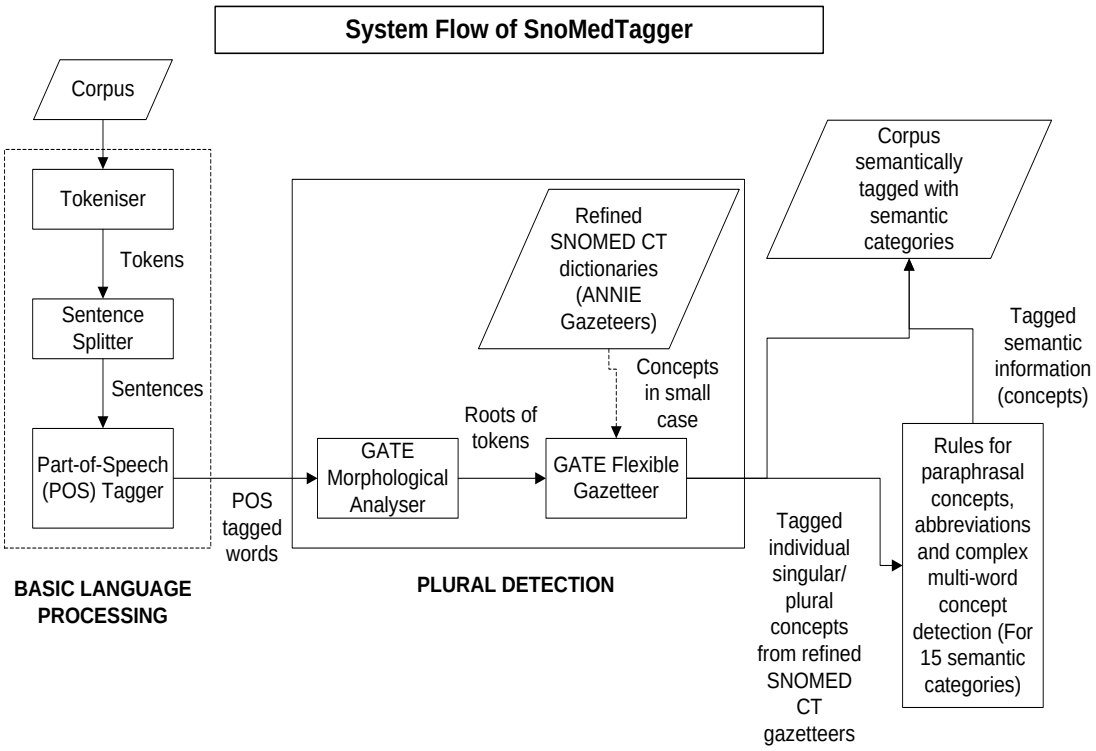


Figure 6-2: System flow of SnoMedTagger.

The application pipeline of SnoMedTagger used 18 CREOLE components (processing resources in GATE tool) out of which 15 were based on JAPE transducers. These 15 JAPE transducers were used for the development of rule-patterns for the 15 semantic categories. The semantic category '*Attribute*' is an exception. This is because the F-measure achieved by the baseline system for this semantic category was high (reported in Table 4-4). In addition, the semantic category 'Attribute' contained only single word concepts after refinement which could be identified using dictionaries.

As shown in Figure 6-2, the first step in the development of the SnoMedTagger was the application of basic language processing resources on the corpus (Development dataset in this case). These resources are defined below.

GATE English Tokeniser – This processing resource joins a JAPE transducer with a normal tokeniser. In addition to the output produced by a normal tokeniser, the JAPE transducer adds rules to identify tokens such as "30s' ", " 'em ", " 's ", " don't ", etc.

Sentence splitter - This processing resource produce the annotation type "Sentence" in corpus by identifying sentence breaks such as a 'full stop'.

Part-of-speech tagger – Part-of-speech tagger is a modified version of the Brill tagger (Brill 1992; Brill 1994; Brill 1995) that assigns part-of-speech tags to each word.

The tokeniser and sentence splitter annotated the corpus with the annotation types 'Token' and 'Sentence', respectively. Then, the part-of-speech tags were assigned as features to each token which were used in the development of rule- patterns.

Following the application of basic language processing resources, 'GATE Morphological analyser' was used prior to the 'Flexible gazetteer' to process the root feature of tokens. This was done for matching plural concepts using the 'Flexible gazetteer'. The 'Flexible gazetteer' provided flexibility to match the customised output and the external gazetteers/dictionaries on the basis of feature used in morphological analysis. In SnoMedTagger, the ANNIE English gazetteer was used as an external gazetteer which contains 16 dictionaries for each semantic category. By external gazetteer we mean that this gazetteer will not be the part of application pipeline, but will serve as an input to flexible gazetteer (an internal gazetteer in the application pipeline).

The flexible gazetteers would first use the output (roots of tokens) produced by morphological analyser and then would use the external gazetteer to identify and classify both types of dictionary concepts (singular and plural) in the corpus. For instance, concept 'legs' in the corpus, the flexible gazetteer will take root 'leg' produced by morphological analyser. Then, this

root will be matched with the singular concept present in the external gazetteer of 'BodyStructure'.

After the identification of the gazetteer concepts in the corpus, a set of generic rule-patterns were developed to be included in the application pipeline of SnoMedTagger.

The rest of this section explains the development of rule-patterns for the identification and classification of paraphrases of concepts, abbreviations of concepts and complex multiword concepts. The generic rule-patterns were derived from two resources.

1. Analysis of language of medical narratives (Development dataset) written by clinicians.
2. Analysis of multiword concepts in the dictionaries of semantic categories (derived from SNOMED CT concept file) during refinement.

The analysis of multiword concepts in the dictionaries of semantic categories was done using the description logic in the SNOMED CT healthcare terminology. Description logic represents the classification of concepts (super concepts and sub concepts) (F. Baader and Nutt. 2002). In SNOMED CT, description logic is meant to define the ontology but is of limited use in identifying the variations of concepts written in medical narratives. Therefore, generic rule-patterns were developed by analysing a real world dataset (Development dataset), and these rule-patterns were analysed during the refinement of concepts containing description logic in the dictionaries. During the refinement process, these rule-patterns were analysed considering their applicability on medical narratives. All rule-patterns were written using JAPE transducers in GATE tool as follows;

Rule-pattern → Rule- action

Here, the left hand side (LHS) consists of the rule-patterns developed for the identification of concepts and the right hand side (RHS) performs classification of the semantic categories on matching (→) rule-patterns. All semantic categories defined in the rule-patterns are italicised in following sections.

6.2.1 Corpus-based rule-patterns and rule-patterns derived during refinement of concepts

As mentioned earlier, the rule-patterns were derived by analysis of two cases; Language of medical narratives (Development dataset) written by clinicians and multiword concepts observed in the dictionaries of semantic categories during refinement. These rule-patterns were developed using refined SNOMED CT dictionaries and the linguistic features that were identified by a part-of-speech (POS) tagger. In the following section, examples of rule-patterns for the cases used for the identification of paraphrases of concepts, abbreviations of concepts, complex multiword concepts and their classification with respective semantic categories, are presented. In the examples of rule-patterns, all the semantic categories are italicised. The other notations used in the examples are as follows;

sp= Space Token excluding newlines and tab spaces.

IN= Preposition or sub coordinating conjunction (category of token)

CC= Coordinating conjunction (category of token)

DT= Determiner (category of token)

|=Or

{Token.kind==punctuation,SpaceToken.string !=~ "[\\n\\r]"} = All punctuation marks by restricting new lines and tab spaces.

{Token.position==startpunct, SpaceToken.string !=~ "[\\n\\r]} = All punctuation marks which indicate starting positions such as “, ‘, (, {, [, etc. where spaces are not equal to new lines or tab spaces.

{Token.position==endpunct, SpaceToken.string !=~ "[\\n\\r]} = All punctuation marks which indicate ending positions such as ”, ’,), },], etc. where spaces are not equal to new lines or tab spaces.

Lookup.majorType = *BodyStructure* (dictionary of individual body structures such as 'chest', 'pelvis', 'leg', 'abdomen', etc.)

Lookup.majorType = *Procedure* (dictionary of individual procedures such as 'X-Ray', 'radiography', 'CT scan', 'biopsy', etc.)

Lookup.majorType = *QualifierValue* (dictionary of individual qualifier values such as 'left', 'right', 'upper', 'lower', etc.)

Lookup.majorType = *Disorder* (dictionary of individual disorders such as 'trauma', 'infection', 'fracture', 'depression', etc.)

Lookup.majorType = *Situation* (dictionary of individual situations such as 'history', 'postoperative', 'preoperative', etc.)

6.2.1.1 Identification and classification of paraphrases of concepts

The paraphrases of concepts that were missed by the SNOMED CT dictionary application were identified during the refinement process and during the analysis of concepts in the Development dataset. To identify paraphrases of a multiword concept, generic rule-patterns were developed. Examples of such generic rule-patterns are as follows.

Examples of corpus-based rule-patterns

Example-1

Concept in the corpus: ‘X-ray of the chest’ – *Procedure*

Possible paraphrases of this concept: ‘Radiography of chest’, ‘Radiography of the chest’, ‘X-ray of chest’, ‘X-Ray of the chest’, ‘Chest X-Ray’, ‘Chest x-ray’, ‘Chest CXR’, and so on.

In above mentioned paraphrases, individual concepts (such as ‘Radiography’, ‘X-ray’, ‘CXR’, etc.) can be identified and classified by dictionaries. For the identification and classification of complete paraphrases, following rule-patterns were developed.

Rule: Procedure

(

{Lookup.majorType = *Procedure*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*} |

{Lookup.majorType = *Procedure*} {sp} {IN} {sp} {DT} {sp} {Lookup.majorType = *BodyStructure*} |

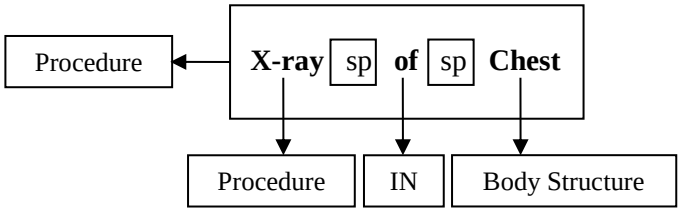
{Lookup.majorType = *BodyStructure*} {sp} {Lookup.majorType = *Procedure*}

): label

→

: label. Procedure = {Rule = Procedure}

For instance, first pattern in this rule can be described as follows;



These rule-patterns are general and will extract other concepts such as; 'GI Prophylaxis', 'pelvic lymphadenectomy', 'abdomen x-ray', 'Prostate biopsy', 'X-Ray of abdomen' and so on.

Example-2

Concept in the corpus: 'CT of the head and neck' – *Procedure*

Possible paraphrases of this concept: 'CT of head and neck', 'CT of the head, neck', 'CT of head, neck', 'CT-Scan of neck and head', 'CT-Scan of head and neck', 'CT scan of the head and neck', 'CT scan of head and neck', 'CT scan of neck and head', 'CT scan of the neck and head', 'CT: head and neck', 'Head CT scan and Neck CT scan', 'CT-Head and Neck', and etc.

Generic rule-patterns for such concepts were written as follows;

Rule: Procedure

(

{Lookup.majorType = *Procedure*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*} {sp} {CC} {sp} {Lookup.majorType = *BodyStructure*} |

{Lookup.majorType = *Procedure*} {sp} {IN} {sp} {DT} {sp} {Lookup.majorType = *BodyStructure*} {sp} {CC} {sp} {Lookup.majorType = *BodyStructure*} |

{Lookup.majorType = *Procedure*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*} {Token.kind = = punctuation, SpaceToken.string !=~ "[\\n\\r]"} {sp} {Lookup.majorType = *BodyStructure*} |

{Lookup.majorType = *Procedure*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*} {Token.kind = = punctuation, SpaceToken.string !=~ "[\\n\\r]"} {Lookup.majorType = *BodyStructure*} |

{Lookup.majorType = *Procedure*} {sp} {IN} {sp} {DT} {sp} {Lookup.majorType = *BodyStructure*} {Token.kind = = punctuation, SpaceToken.string !=~ "[\\n\\r]"} {sp} {Lookup.majorType = *BodyStructure*} |

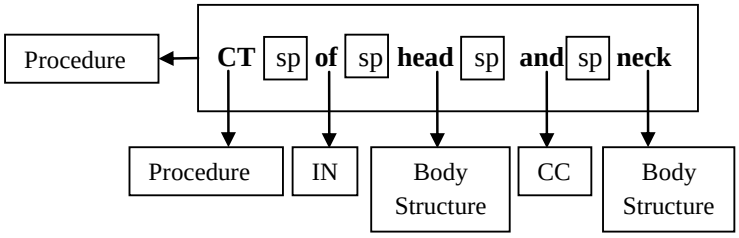
{Lookup.majorType = *Procedure*} {Token.kind = = punctuation, SpaceToken.string !=~ "[\\n\\r]"} {sp} {Lookup.majorType = *BodyStructure*} {sp} {CC} {sp} {Lookup.majorType = *BodyStructure*}

): label

→

: label. Procedure = {Rule = Procedure}

For instance, the first rule-pattern in the above rule will identify paraphrases such as 'CT of head and neck', 'CT of neck and head', 'CT Scan of head and neck', 'CT Scan of neck and head', 'CT-Scan of head and neck', 'CT-Scan of neck and head', 'Ct Scan of head and neck', 'Ct Scan of neck and head', 'ct scan of head and neck', 'ct scan of neck and head', 'Ct-Scan of head and neck', 'Ct-Scan of neck and head', 'Ct-scan of head and neck', 'Ct-scan of neck and head', etc. This rule-pattern is described as follows;



Example-3

Concept in the corpus: 'History of breast cancer' – *Situation*

Possible paraphrases of this concept: 'Breast cancer history', 'history of cancer in breast', 'History: Breast cancer', 'History of cancer of breast', etc.

Generic rule-patterns for such concepts were written as follows;

Rule: Situation

(

{Lookup.majorType = *Situation*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*} {sp} {Lookup.majorType = *Disorder*} |

{Lookup.majorType = *BodyStructure*} {sp} {Lookup.majorType = *Disorder*} {sp} {Lookup.majorType = *Situation*} |

{Lookup.majorType = *Situation*} {Token.kind = = punctuation, SpaceToken.string !=~ "[\\n\\r]"} {sp} {Lookup.majorType = *Disorder*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*} |

{Lookup.majorType = *Situation*} {sp} {IN} {sp} {Lookup.majorType = *Disorder*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*}

): label

→

: label. Situation = {Rule = Situation}

These generic rule-patterns are able to identify and classify other concepts paraphrases of '*Situation*' associated with '*Disorder*' (such as 'History of trauma', 'History of hepatitis B', etc.)

Examples of rule-patterns analysed during refinement of dictionaries

During the refinement process, generic rule-patterns were developed for the identification and classification of paraphrases for multiword concepts in the dictionaries. This identification will not only ensure the applicability of rule-patterns in medical narratives but will also maintain the identification of structured concepts present in original SNOMED CT dictionaries.

Example-1

SNOMED CT concept: 'Artery of thorax and/or abdomen' – *BodyStructure*

Possible paraphrases of this concept: 'Artery of thorax or/and abdomen', 'Thorax artery and abdomen artery', 'Thorax artery or abdomen artery', 'Artery of thorax and abdomen', 'Artery of thorax or abdomen', 'Artery of thorax and artery of abdomen', 'Artery of thorax or artery of

abdomen', 'Artery of abdomen and artery of thorax', etc. Generic rule-patterns for this concept are as follows.

Rule: BodyStructure

(

{Lookup.majorType = *BodyStructure*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*} {sp} {CC} {sp} {Lookup.majorType = *BodyStructure*} |

{Lookup.majorType = *BodyStructure*} {sp} {Lookup.majorType = *BodyStructure*} {sp} {CC} {sp} {Lookup.majorType = *BodyStructure*} {sp} {Lookup.majorType = *BodyStructure*} |

{Lookup.majorType = *BodyStructure*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*} {sp} {CC} {sp} {Lookup.majorType = *BodyStructure*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*}|

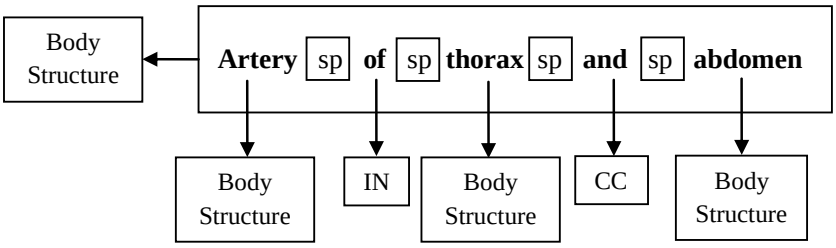
{Lookup.majorType = *BodyStructure*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*} {sp} {CC}{Token.kind = = punctuation, SpaceToken.string !=~ "[\\n\\r]"} {CC}{sp} {Lookup.majorType = *BodyStructure*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*}

): label

→

:label. BodyStructure = {Rule = BodyStructure}

The first rule-pattern is described as follows.



Example-2

SNOMED CT concept: 'Allergic reaction to flour dust' – *Disorder*

Paraphrases of this concept: 'Allergic reaction from flour dust', 'Flour dust allergy', 'Allergic reaction by flour dust', 'Allergic reaction caused by flour dust', etc. Generic rule-patterns for this concept which can also identify and classify any 'Disorder' caused by any 'Product or Substance' are as follows.

Rule: Disorder

(

{Lookup.majorType = *Disorder*}{sp}{Lookup.majorType = *Qualifier_Value*} {sp} {IN} {sp} {Lookup.majorType = *ProductorSubstance*} |

{Lookup.majorType = *ProductorSubstance*} {sp} {Lookup.majorType = *Disorder*}{sp}{Lookup.majorType = *Qualifier_Value*} |

{Lookup.majorType = *Disorder*} {sp} {Lookup.majorType = *Attribute*} {sp} {Lookup.majorType = *ProductorSubstance*}

): label

→

: label. Disorder = {Rule = Disorder}

Example-3

SNOMED CT concept: 'Health clinic managed by voluntary or private agent' – *Environment*

Paraphrases of this concept: 'Health clinic managed by voluntary agent or private agent', 'Health clinic organised by voluntary agent or private agent', 'Health clinic maintained by voluntary agent or private agent', etc. All the mentioned paraphrases can be identified by following rule-patterns.

Rule: Environment

(

{Lookup.majorType = *Environment*} {sp} {Lookup.majorType = *Attribute*} {sp} {Lookup.majorType = *QualifierValue*} {sp} {Lookup.majorType = *Occupation*} {sp} {CC} {sp} {Lookup.majorType = *QualifierValue*} {Lookup.majorType = *Occupation*}

): label

→

: label. Environment = {Rule = Environment}

6.2.1.2 Identification and classification of abbreviation of concepts

As mentioned earlier in Section 6.1.2.2, clinicians prefer to write either abbreviations of concepts or the definitions of a concept, but not both. The multiword concepts containing abbreviations of concepts were analysed in the Development dataset. During the refinement process, the abbreviation of concepts were separated from its definition and stored in the

relevant dictionary of a semantic category. Therefore, a multiword concept and/or its paraphrase containing abbreviation can be identified and classified by the generic rule-patterns. Examples of such rule-patterns are as follows.

Example of corpus-based rule-pattern

Concepts in the corpus such as 'GI prophylaxis', 'Chest CXR', 'lung CXR', 'Ct abdomen', etc., the individual concepts can be identified by dictionaries while the multiword concepts containing abbreviations can be identified by following rule-pattern.

Rule: Procedure
(
{Lookup.majorType = *BodyStructure*} {sp} {Lookup.majorType = *Procedure*}
): label
→
:label.Procedure = {Rule = Procedure}

Examples of rule-patterns analysed during refinement of dictionaries

Although before refinement the multiword concepts contained abbreviation with their definitions which were not present in the medical narratives, generic rule-patterns were still developed in order to identify abbreviations with their definitions that are commonly present in a structured dataset (such as MEDLINE abstracts). This was done so that the semantic tagger would not lose semantic information if it was applied on a structured dataset and to maintain the structure of the SNOMED CT dictionary concepts.

Examples

SNOMED CT concept: DVT - Deep venous thrombosis – *Disorder* can be written in several other forms;

Paraphrases of this concept: DVT, Deep venous thrombosis, DVT (Deep venous thrombosis), (Deep venous thrombosis), DVT, Deep venous thrombosis, (Deep venous thrombosis) DVT,

DVT (Deep venous thrombosis), (DVT), DVT: Deep venous thrombosis, Deep venous thrombosis: DVT, etc.

The generic rule-patterns developed are as follows.

Rule: Disorder

(

{Lookup.majorType = *Disorder*} {sp} {Token.kind = = punctuation, SpaceToken.string !=~ "[\\n\\r]"} {sp} {Lookup.majorType = *Disorder*} |

{Lookup.majorType = *Disorder*} {Token.kind = = punctuation, SpaceToken.string !=~ "[\\n\\r]"} {sp} {Lookup.majorType = *Disorder*} |

{Lookup.majorType = *Disorder*} {Token.kind = = punctuation, SpaceToken.string !=~ "[\\n\\r]"} {Lookup.majorType = *Disorder*} |

{Token.kind = = punctuation, Token.position = = startpunct, SpaceToken.string !=~ "[\\n\\r]"} {Lookup.majorType = *Disorder*} {Token.kind = = punctuation, Token.position = = endpunct, SpaceToken.string !=~ "[\\n\\r]"} |

{Token.kind = = punctuation, Token.position = = startpunct, SpaceToken.string !=~ "[\\n\\r]"} {Lookup.majorType = *Disorder*} {Token.kind = = punctuation, Token.position = = endpunct, SpaceToken.string !=~ "[\\n\\r]"} {sp}{Lookup.majorType = *Disorder*}|

{Token.kind = = punctuation, Token.position = = startpunct, SpaceToken.string !=~ "[\\n\\r]"} {Lookup.majorType = *Disorder*} {Token.kind = = punctuation, Token.position = = endpunct, SpaceToken.string !=~ "[\\n\\r]"} {Lookup.majorType = *Disorder*}|

{Lookup.majorType = *Disorder*} {sp} {Token.kind = = punctuation, Token.position = = startpunct, SpaceToken.string !=~ "[\\n\\r]"} {Lookup.majorType = *Disorder*} {Token.kind = = punctuation, Token.position = = endpunct, SpaceToken.string !=~ "[\\n\\r]"} |

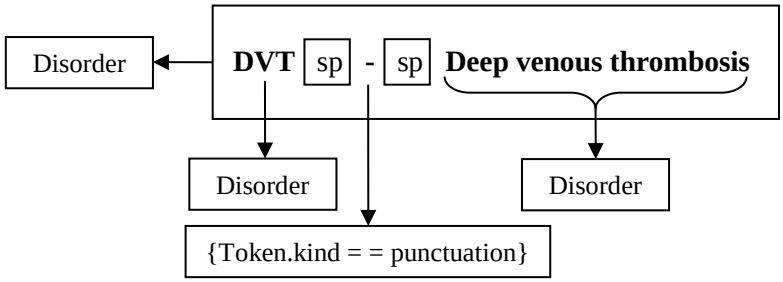
{Lookup.majorType = *Disorder*} {Token.kind = = punctuation, Token.position = = startpunct, SpaceToken.string !=~ "[\\n\\r]"} {Lookup.majorType = *Disorder*} {Token.kind = = punctuation, Token.position = = endpunct, SpaceToken.string !=~ "[\\n\\r]"}

): label

→

: label.Disorder = {Rule = Disorder}

First rule-pattern is described as follows.



Similar rule-patterns were developed for the other semantic categories containing abbreviations and their definitions. Examples of such cases include Liver function test – LFT, IV – Intravenous, RBC – Red blood cell, WBC – White blood cell, etc.

6.2.1.3 Identification and classification of complex multiword concepts

Clinicians write complex multiword concepts in medical narratives which dictionaries find difficult to identify. These multiword concepts are complex because these concepts contain overlapped concepts associated with more than one semantic category. All such complex multiword concepts require rule-patterns for all overlapped semantic categories. Examples of generic rule-patterns developed for complex multiword concepts are as follows.

Examples of corpus-based rule-patterns

Example-1

Concept in the corpus: 'Bilateral pelvic lymph node dissection' – *Procedure*

Individual concepts such as 'bilateral – *QualifierValue*', 'pelvic – *BodyStructure*', 'lymph node – *BodyStructure*', 'dissection – *Procedure*' can be identified by the relevant dictionaries. For the complex multiword concepts related to 'Procedure' and 'BodyStructure', the following generic rule-patterns were developed.

Rule: Procedure

(

{Lookup.majorType = *QualifierValue*} {sp} {Lookup.majorType = *BodyStructure* }{sp}{Lookup.majorType = *BodyStructure* }{sp} {Lookup.majorType = *Procedure*} |

{Lookup.majorType = *BodyStructure* }{sp}{Lookup.majorType = *BodyStructure* }{sp} {Lookup.majorType = *Procedure*}|

{Lookup.majorType = *BodyStructure* }{sp} {Lookup.majorType = *Procedure*}

): label

→

:label.Procedure = {Rule = Procedure}

Rule: BodyStructure

(

{Lookup.majorType = *QualifierValue*} {sp} {Lookup.majorType = *BodyStructure* }{sp}{Lookup.majorType = *BodyStructure* } |

{Lookup.majorType = *BodyStructure* }{sp}{Lookup.majorType = *BodyStructure* }

): label

→

:label.BodyStructure = {Rule = BodyStructure}

Example-2

Concept in the corpus: 'History of autoimmune hepatitis with cirrhosis' – *Situation*

Individual concepts such as 'History – *Situation*', 'autoimmune – *Disorder*', 'hepatitis – *Disorder*', 'cirrhosis – *Disorder*' can be identified by the relevant dictionaries and complex multiword concepts related to 'Situation' and 'Disorder' will be identified by the following generic rule- patterns.

Rule: Situation

(

{Lookup.majorType = *Situation*} {sp} {IN} {sp} {Lookup.majorType = *Disorder*} |

{Lookup.majorType = *Situation*} {sp} {IN} {sp} {Lookup.majorType = *Disorder*} {sp} {IN} {sp} {Lookup.majorType = *Disorder*}|

{Lookup.majorType = *Situation*} {sp} {IN} {sp} {Lookup.majorType = *Disorder*}{sp} {Lookup.majorType = *Disorder*} {sp} {IN} {sp} {Lookup.majorType = *Disorder*}

): label

→

:label.Situation = {Rule = Situation}

Rule: Disorder

(

{Lookup.majorType = *Disorder*} {sp} {Lookup.majorType = *Disorder*}|

{Lookup.majorType = *Disorder*} {sp} {IN} {sp} {Lookup.majorType = *Disorder*}|

{Lookup.majorType = *Disorder*}{sp} {Lookup.majorType = *Disorder*} {sp} {IN} {sp} {Lookup.majorType = *Disorder*}

): label

→

:label.Disorder = {Rule = Disorder}

Example-3

Concept in the corpus: 'Cesarean section for rupture of membranes' – *Procedure*

Individual concepts such as 'Cesarean section – *Procedure*', 'rupture – *Disorder*', 'membranes – *BodyStructure*' can be identified by the relevant dictionaries and complex multiword concepts related to 'Procedure' and 'Disorder' can be identified by the following generic rule-patterns.

Rule: Procedure

(

{Lookup.majorType = *Procedure*} {sp} {IN} {sp} {Lookup.majorType = *Disorder*}{sp} {IN} {sp} {Lookup.majorType = *BodyStructure*}

): label

→

:label.Procedure = {Rule = Procedure}

Rule: Disorder

(

{Lookup.majorType = *Disorder*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*}

): label

→

:label.Disorder = {Rule = Disorder}

Examples of rule-patterns analysed during refinement of dictionaries

Example-1

SNOMED CT concept: 'Excision of birthmark of head or neck' – *Procedure*

Individual concepts such as ‘Excision – *Procedure*’, ‘birthmark – *Disorder*’, ‘head – *BodyStructure*’, ‘neck – *BodyStructure*’ can be identified by the relevant dictionaries and complex multiword concepts related to ‘Procedure’, ‘BodyStructure’ and ‘Disorder’ can be identified by following the generic rule- patterns.

Rule: Procedure

(

{Lookup.majorType = *Procedure*} {sp} {IN} {sp} {Lookup.majorType = *Disorder*} |

{Lookup.majorType = *Situation*} {sp} {IN} {sp} {Lookup.majorType = *Disorder*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*}

): label

→

:label.Procedure = {Rule = Procedure}

Rule: BodyStructure

(

{Lookup.majorType = *BodyStructure*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*}

): label

→

:label. BodyStructure = {Rule = BodyStructure}

Rule: Disorder

(

{Lookup.majorType = *Disorder*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*} |

{Lookup.majorType = *Disorder*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*} {sp} {CC} {sp} {Lookup.majorType = *BodyStructure*}

): label

→

:label.Disorder = {Rule = Disorder}

Example-2

SNOMED CT concept: ‘Erythema of mucous membrane of mouth’ – *Findings*

Individual concepts such as ‘Erythema – *Findings*’, ‘mucous – *QualifierValue*’, ‘membrane – *BodyStructure*’, ‘mouth – *BodyStructure*’ can be identified by the relevant dictionaries and complex multiword concepts related to ‘Findings’ and ‘BodyStructure’ can be identified by the following generic rule- patterns:

Rule: Findings

(

{Lookup.majorType = *Findings*} {sp} {IN} {sp} {Lookup.majorType = *QualifierValue*}{sp} {Lookup.majorType = *BodyStructure*} |

{Lookup.majorType = *Findings*} {sp} {IN} {sp} {Lookup.majorType = *QualifierValue*}{sp} {Lookup.majorType = *BodyStructure*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*}

): label

→

:label.Findings = {Rule = Findings}

Rule: BodyStructure

(

{Lookup.majorType = *QualifierValue*}{sp} {Lookup.majorType = *BodyStructure*} |

{Lookup.majorType = *QualifierValue*}{sp} {Lookup.majorType = *BodyStructure*} {sp} {IN} {sp} {Lookup.majorType = *BodyStructure*}

): label

→

:label.Findings = {Rule = Findings}

Similarly, N=316 generic rule-patterns were written for the 15 semantic categories by analysing all possible combinations of refined dictionaries and linguistic features, as shown in Table 6-2. The performance measurements achieved by ‘SnoMedTagger’ on the Development dataset are presented in Table 6-3.

After achieving better performance measurements on the Development dataset against the baseline results, SnoMedTagger was evaluated and validated using two different unseen gold standard Test datasets (Chapter 7).

Table 6-2: Successful combinations of refined dictionaries and linguistic features used in the development of rule-patterns for SnoMedTagger.

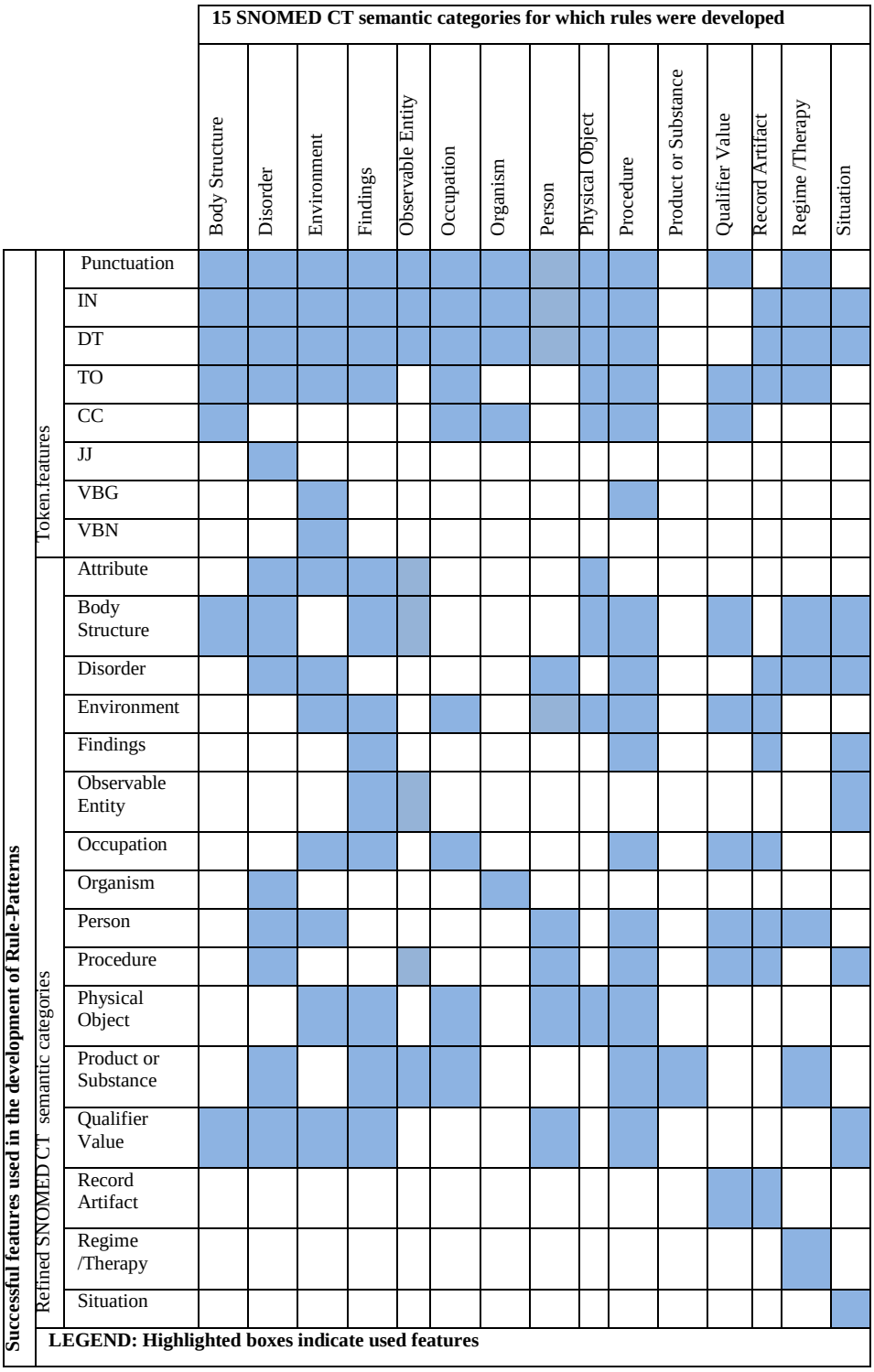


Table 6-3: Performance measurements achieved by SnoMedTagger on Development dataset.

Semantic Categories	True positives	Actual concepts in Gold Standard	Recall (%)	Precision (%)	F-measure (%)
Attribute	589	640	92	92	92
BodyStructure	156	221	71	88	78
Disorder	291	376	77	78	77
Environment	216	226	96	82	89
Findings	347	446	78	75	76
ObservableEntity	129	164	79	72	75
Occupation	81	94	86	70	77
Organism	5	7	71	100	83
Person	173	203	85	98	91
PhysicalObject	80	114	70	73	72
Procedure	532	697	76	77	76
ProductorSubstance	272	385	71	93	80
QualifierValue	996	1347	74	76	75
Record Artifact	41	42	98	93	95
Regime/Therapy	74	102	73	86	79
Situation	45	61	74	74	74
Micro summary	4027	**5125**	79	81	80

6.3 Summary

In this chapter, the resources and tools that were used in this research were described and the process of the refinement of the dictionaries was explained in detail. The refined dictionaries were then used along with linguistic features for the development of SnoMedTagger.

The complete system flow of SnoMedTagger was described and the development of generic rule-patterns for the identification of semantic information (paraphrases of concepts, abbreviations of concepts and complex multiword) and their classification with semantic categories were presented. These generic rule-patterns were analysed using two resources; Medical narratives (the Development dataset) written by clinicians, and multiword concepts in the dictionaries of the semantic categories during refinement. Examples of rule-patterns analysed for both resources were also discussed in this chapter. The next chapter contains a complete evaluation and validation of the SnoMedTagger developed in this research.

Chapter 7. Evaluation and validation of the SnoMedTagger

7.1 Introduction

In Computational Linguistics, the performance of natural language processing applications (such as named entity recognition systems and semantic taggers) is usually evaluated against human-annotated gold standards. Following this approach, the SnoMedTagger, which was developed on the basis of rule-patterns, was evaluated against two human annotated gold standard test datasets (described in Chapter 5). In addition, the performance of SnoMedTagger was also compared with the performance of the following systems.

1. Baseline system (SNOMED CT dictionary application).
2. SVM-based machine learning system.
3. Ontology-based BioPortal web annotator.

The evaluation using the two datasets proved that the SnoMedTagger can be applied on datasets that were quite different in origin and nature. Furthermore, the comparison with the other systems reviewed its performance against other approaches/methods. After these evaluations, the semantic information that was identified by SnoMedTagger was also validated by two domain experts. This validation identified changes which have helped in improving the refined dictionaries used by the SnoMedTagger.

The following sections describe in detail the evaluation and validation that was carried out. First, the performance of the baseline system, the SnoMedTagger, the BioPortal web annotator and an SVM-based system was evaluated against the two gold standard test datasets; then, the performance of SnoMedTagger was compared with the performance of the other three systems. Lastly, the semantic information in the form of the output concepts was validated by two domain experts who were not the original annotators. To the best of our knowledge, SnoMedTagger is the first semantic tagger for medical narratives that has been developed using

globally known semantic categories derived from SNOMED CT (Hina, Atwell and Johnson 2013a; Hina, Atwell and Johnson 2013b) and methodologically evaluated and validated on more than one human-annotated datasets.

The standard metrics of recall, precision and f-measure were used in all the evaluations. These metrics are defined as follows.

Recall is the percentage measurement that shows the number of correctly identified terms. The higher the recall rate, the better the system is in identifying correct terms. Terms represent concepts in this research. Recall is calculated using this formula;

$$Recall = \frac{tp}{tp + fn}$$

Here,

tp = True positives: Correct concepts that should be identified.

fn= False negatives: Concepts that should match but did not match by the application.

Precision is the percentage measurement that shows the number of identified terms (concepts) regardless of whether the system failed to retrieve correct terms. The formula used for calculating precision is as follows;

$$Precision = \frac{tp}{tp + fp}$$

Here,

fp = False positives: Concepts that matched by the application but should not be identified.

'F-measure' is the percentage measurement that shows the trade-off between 'Precision' and 'Recall'. This was calculated using following formula;

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

The following analyses are based on the recall, precision and f-measure of the individual semantic categories. In addition, a micro summary which provides the recall, precision and f-measure of the whole corpus, was used to compare the different systems.

7.2 Evaluation of baseline system using gold standard test datasets

As described in Section 4.3, a baseline system was developed using the dictionaries of semantic categories that were derived from SNOMED CT. This baseline system was tested on gold standard test datasets annotated by domain experts that were described in Chapter 5. The baseline results were used to monitor the performance of SnoMedTagger during development and for comparison against other methods.

The baseline results for the Test dataset 1 are tabulated in Table 7-1. These results indicate that the baseline system did not perform satisfactorily on the individual semantic categories. The semantic categories 'Attribute' and 'Person' are an exception.

Table 7-1: Evaluation of baseline system against gold standard Test dataset 1.

Semantic categories	**True positives**	**Actual concepts in Gold Standard**	**Recall (%)**	**Precision (%)**	**F-measure (%)**
Attribute	285	299	**95**	**74**	**83**
BodyStructure	4	25	15	40	22
Disorder	30	167	19	86	31
Environment	24	74	48	96	64
Findings	108	244	46	78	58
ObservableEntity	129	171	75	69	72
Occupation	16	68	26	67	37
Organism	**Not present**				
Person	164	200	**83**	**98**	**90**
PhysicalObject	0	40	0	0	0
Procedure	39	125	34	83	48
ProductorSubstance	134	222	64	58	61
QualifierValue	492	848	58	52	55
RecordArtifact	0	17	0	0	0
Regime/Therapy	23	96	28	70	40
Situation	37	76	51	86	64
Micro summary	1485	**2672**	58	65	62

Most of the concepts in the semantic category 'Attribute' were single word concepts. As a result of this, the baseline system, which is based on dictionaries, identified a large number of 'Attribute' concepts in the Test dataset 1. In case of the semantic category 'Person', 164 out of 200 concepts were correctly identified by the baseline system, thus resulting in a high f-measure value of 90%. This can be attributed to the fact that only a few relevant multiword concepts were present in the Test dataset 1. The concepts associated with the semantic category 'Organism' [9] were not present in the Test dataset 1. Thus, the performance metrics for this semantic category could not be calculated for this dataset.

The results of the evaluation of the baseline system, using Test dataset 2, are provided in Table 7-2. As in the case of evaluation using Test dataset 1, the high f-measure for the semantic category 'Attribute' was related to the fact that the Test dataset 2 also contained a large number of single word concepts associated with this semantic category.

Table 7-2: Evaluation of baseline system against gold standard Test dataset 2.

Semantic categories	**True positives**	**Actual concepts in Gold Standard**	**Recall (%)**	**Precision (%)**	**F-measure (%)**
Attribute	1369	1500	**91**	**77**	**83**
BodyStructure	195	1445	13	86	23
Disorder	693	1442	48	94	64
Environment	268	532	50	87	64
Findings	715	2958	24	80	37
ObservableEntity	202	1195	17	68	27
Occupation	34	222	15	87	26
Organism	**20**	**57**	35	61	44
Person	505	654	**77**	**99**	**87**
PhysicalObject	47	630	7	72	14
Procedure	630	2300	27	96	43
ProductorSubstance	864	1619	53	86	66
QualifierValue	3583	5440	66	57	61
RecordArtifact	56	130	43	100	60
Regime/Therapy	32	408	8	64	14
Situation	54	321	17	42	24
Micro Summary	9267	**20853**	44	71	55

[9] False positives for the semantic category 'Organism' were not identified.

Similarly, all the single word ‘Person’ concepts in the Test dataset 2 were identified by the baseline system. However, the multiword concepts were left unidentified, resulting in lowering the recall rate to 77% for this semantic category. The evaluation of the baseline system for the semantic category ‘Organism’ was done using Test dataset 2. Out of 57 ‘Organism’ concepts that were present in the Test dataset 2, only 20 concepts were identified by the baseline system.

The evaluation presented above indicated clearly that the baseline system did not perform satisfactorily for almost all of the semantic categories in the two test datasets. Thus, it was concluded that the use of dictionary-based approach is not appropriate for identification of semantic information in medical narratives.

7.3 Evaluation of SnoMedTagger using the gold standard test datasets

As described in Chapter 6, the SnoMedTagger was developed by considering the language issues (Table 4-5) in the concepts that were not identified by the baseline system. The SnoMedTagger was then evaluated using the gold standard test datasets. The results of these evaluations are discussed in the following text. The results of evaluation of the SnoMedTagger, using Test dataset 1, are presented in Table 7-3. These results clearly indicate that the SnoMedTagger performed better than the baseline system. This is depicted in Figure 7-1. Generally, the recall, precision and f-measure were considerably greater compared to those achieved by the baseline system. This can be attributed to the superior capability of the developed rule patterns to identify semantic information.

For instance, the baseline system did not identify any of the concepts in the semantic categories ‘Physical Object’ and ‘Record Artifact’. In contrast, the performance the SnoMedTagger was remarkably high as indicated by f-measure score of 84% and 65% for these two semantic categories, respectively.

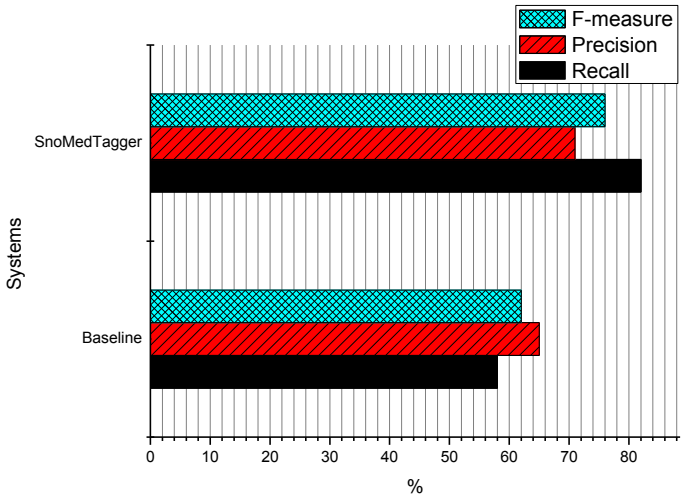


Figure 7-1: Overall comparison of the baseline system and the SnoMedTagger on Test dataset 1.

Similarly, the rule-patterns that were developed resulted in the correct identification of multiword concepts in the semantic category 'Person'. As a result, the f-measure achieved by the SnoMedTagger for this category was 8% higher compared to that achieved by the baseline system.

Table 7-3: Evaluation of SnoMedTagger against gold standard Test dataset 1.

Semantic Categories	**True positives**	**Actual concepts in Gold Standard**	**Recall (%)**	**Precision (%)**	**F-measure (%)**
Attribute	277	299	**93**	**80**	**86**
BodyStructure	24	25	92	65	76
Disorder	125	167	75	81	78
Environment	48	74	65	87	74
Findings	171	244	70	59	64
ObservableEntity	120	171	70	70	70
Occupation	59	68	87	73	79
Organism	**Not present**				
Person	196	200	**98**	**98**	**98**
PhysicalObject	38	40	95	75	**84**
Procedure	108	125	86	50	63
ProductorSubstance	191	222	86	79	83
QualifierValue	693	848	82	65	72
RecordArtifact	10	17	59	71	65
Regime/Therapy	81	96	84	78	**81**
Situation	60	76	79	79	79
Micro summary	2201	**2672**	82	71	76

The results of evaluation of the SnoMedTagger, using Test dataset 2, are presented in Table 7-4. The results clearly indicate that the SnoMedTagger performed considerably better than the baseline system, as depicted in Figure 7-2.

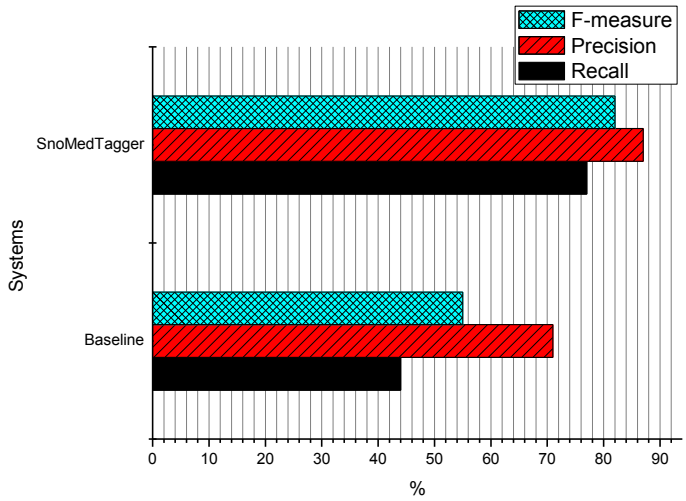


Figure 7-2: Overall comparison of the baseline system and the SnoMedTagger on Test dataset 2.

Table 7-4: Evaluation of SnoMedTagger against gold standard Test dataset 2.

Semantic categories	**True positives**	**Actual concepts in Gold Standard**	**Recall (%)**	**Precision (%)**	**F-measure (%)**
Attribute	1480	1500	99	85	91
BodyStructure	1169	1445	81	97	88
Disorder	1062	1442	74	98	84
Environment	420	532	79	94	86
Findings	1461	2958	49	84	**62**
ObservableEntity	370	1195	31	81	**45**
Occupation	138	222	62	77	69
Organism	25	57	**44**	**34**	**38**
Person	587	654	90	99	94
PhysicalObject	482	630	77	81	79
Procedure	1841	2300	80	96	87
ProductorSubstance	1134	1619	70	98	82
QualifierValue	5379	5440	99	85	91
RecordArtifact	124	130	95	87	91
Regime/Therapy	75	408	18	88	**30**
Situation	233	321	73	46	**56**
Micro Summary	15980	**20853**	77	87	82

It can be seen from the results that the SnoMedTagger achieved low f-measures for the semantic categories such as ‘Findings’, ‘Observable Entity’, ‘Regime/Therapy’ and ‘Situation’. This was due to the fact that the Test dataset 2 was different in terms of content from Test dataset 1. In addition to this fact, Test dataset 2 contained large number of concepts for each individual semantic category than the Test dataset 1.

The semantic category ‘Organism’ is also an exception. For this semantic category, the SnoMedTagger achieved f-measure of 38% while the baseline system achieved f-measure of 44%. However, it was noted that the recall score achieved by the SnoMedTagger was 9% higher than that achieved by the baseline system. This is because of the fact that the Development dataset contained only 7 ‘Organism’ concepts out of which 3 were identified by the baseline system. Thus, we were unable to develop the rule-patterns on the basis of the analysis of an appropriately large number of ‘Organism’ concepts.

In summary, the evaluation of the SnoMedTagger presented above indicates that the developed rule-patterns can be applied on different datasets (medical narratives) for the identification of semantic information. For the comparison of rule-based approach of SnoMedTagger with other approaches, ontology-based BioPortal web annotator and SVM-based machine learning system was tested using same gold standard test datasets, explained in the next sections.

7.4 Evaluation of BioPortal web annotator using gold standard test datasets

‘BioPortal’ is a web portal which provides a selection of over 300 ontologies from the biological and medical domain (Noy et al. 2009). In this research, Bioportal web annotator was employed to annotate the test datasets using the SNOMED CT ontology. The Bioportal web annotator provides Python client code for the annotation of large datasets. This code was used to annotate the concepts in Test dataset 1 and Test dataset 2 with the selected 16 SNOMED CT semantic categories. These annotations were then compared against the two human-annotated gold standard test datasets.

The results of evaluation of the Bioportal web annotator, using Test dataset 1, are tabulated in Table 7-5. In comparison to the baseline system and the SnoMedTagger, the performance of Bioportal web annotator was inferior in case of the semantic category 'Attribute'. This is depicted in Figure 7-3. This is because a number of linkage-type 'Attribute' concepts identified by domain experts in the Test dataset 1 were not identified by the Bioportal web annotator. Examples of such concepts are 'with', 'after', 'in' and so in. This is because the BioPortal web annotator considered such concepts as stop words in the input text.

Table 7-5: Evaluation of BioPortal web annotator against gold standard Test dataset 1.

Semantic Categories	**True positives**	**Actual concepts in Gold Standard**	**Recall (%)**	**Precision (%)**	**F-measure (%)**
Attribute	74	299	**25**	**43**	**31**
BodyStructure	14	25	82	30	44
Disorder	102	167	64	39	48
Environment	24	74	35	86	50
Findings	117	244	48	34	40
ObservableEntity	89	171	53	26	35
Occupation	55	68	89	62	73
Organism	**Not present**				
Person	160	200	81	61	70
PhysicalObject	3	40	10	13	12
Procedure	48	125	41	36	38
ProductorSubstance	183	222	88	48	62
QualifierValue	389	848	46	36	41
RecordArtifact	2	17	25	100	40
Regime/Therapy	17	96	21	52	30
Situation	48	76	67	72	69
Micro summary	1325	**2672**	52	40	45

Furthermore, it can be said, on the basis of the performance metrics provided in Table 7-5, that the BioPortal web annotator did not perform satisfactorily for any of the other semantic categories, generally. This was because of the limited language of concepts in the SNOMED CT ontology. Thus the ontology can be regarded as inappropriate to deal with the variation in writing styles found in medical narratives. This point is also considered in Section 5.3.1.

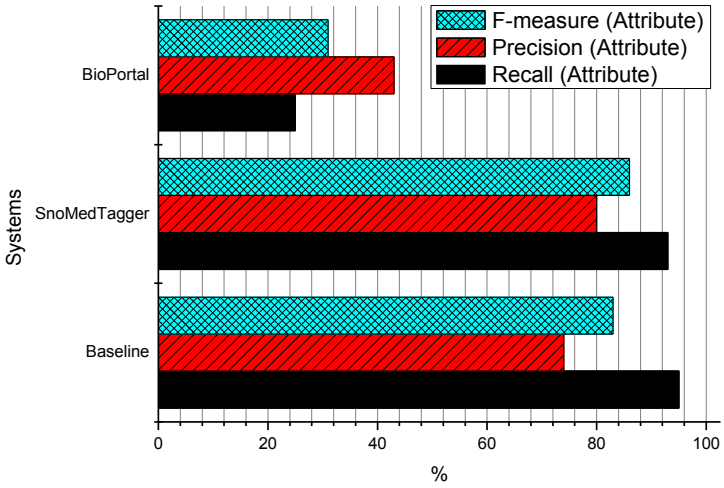


Figure 7-3: Comparison of baseline system, Bioportal web annotator and SnoMedTagger for the semantic category 'Attribute' in Test dataset 1.

In case of the Test dataset 2, the generally low performance of the Bioportal web annotator is evident from the data provided in Table 7-6. Again, this low performance was due to the limited language of concepts in the SNOMED CT ontology.

Table 7-6: Evaluation of BioPortal web annotator against gold standard Test dataset 2.

Semantic categories	**True positives**	**Actual concepts in Gold Standard**	**Recall (%)**	**Precision (%)**	**F-measure (%)**
Attribute	544	1500	35	64	45
BodyStructure	863	1445	62	52	57
Disorder	893	1442	64	49	56
Environment	303	532	59	54	56
Findings	959	2958	33	27	30
ObservableEntity	240	1195	22	43	29
Occupation	62	222	28	81	42
Organism	39	57	**75**	**29**	**41**
Person	506	654	77	75	76
PhysicalObject	144	630	24	69	35
Procedure	779	2300	35	49	41
ProductorSubstance	900	1619	60	36	45
QualifierValue	3487	5440	62	33	43
RecordArtifact	29	130	22	94	36
Regime/Therapy	22	408	6	35	11
Situation	80	321	23	48	31
Micro Summary	9850	**20853**	48	39	43

In comparison to the baseline system and the SnoMedTagger, the BioPortal achieved better recall rate (75%) in case of the semantic category ‘Organism’. This is presented in Figure 7-4.

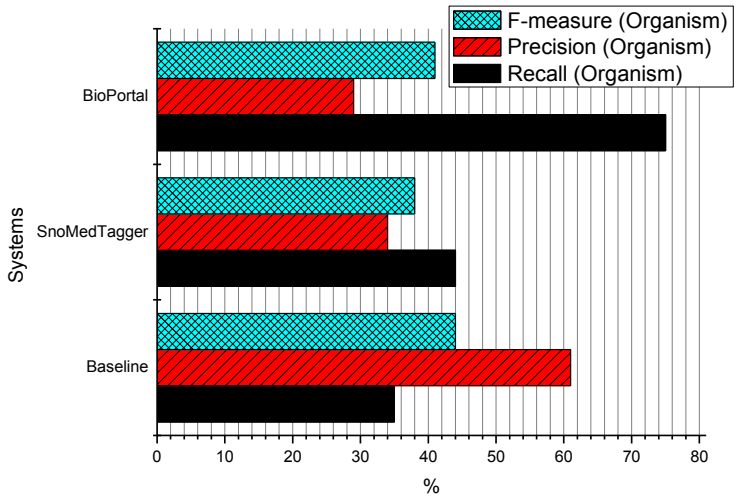


Figure 7-4: Comparison of baseline system, BioPortal web annotator and SnoMedTagger for the semantic category 'Organism' in Test dataset 2.

This is because the concepts in this semantic category are mostly proper names. Thus, in a number of cases, the exact names are present in both the ‘concept table’ and ‘description table’ in the SNOMED CT ontology. As a result, the BioPortal identified such concepts more than once, resulting in an apparently high recall rate. The low precision for this semantic category was because of the fact that a large number of false positives were identified in the Test dataset 2.

7.5 Evaluation of an SVM-based machine learning system using gold standard test datasets

To compare the performance of the baseline system and the SnoMedTagger with the performance of an appropriate machine learning approach, LibSVM, which was available as GATE tool, was used. LibSVM is a Java version of the Support Vector Machines (SVMs) package (Li and Shawe-Taylor 2003). In language processing, SVM is well-known for classification tasks. The system is capable of learning features with high generalisation using a

Kernel function. In the present study, we used a linear Kernel with the extension of multiple classifications (‘one vs. others’).

In case of machine learning, a large annotated data of similar nature is preferred for training and testing. This is because such a data facilitates the system in learning the more appropriate generic features. However, in this study, we were unable to do so due to resource constraints such as the cost of annotating a large dataset. Therefore, for training of the SVM classifier, we used the Development dataset that contained 5125 concepts. This was considered a reasonable number of concepts for the overall training but the number of concepts associated with individual semantic categories was limited. The general feature set, which was used in the development of rule-patterns, was also used to train the classifier on the Development dataset.

Multiple ranges and different features were tested and the best performing features and ranges, which are listed below, were used in the training task.

1. Refined SNOMED CT dictionaries (for chunking individual concepts).

2. Part-of-speech categories of three words before and three words after the dictionary concepts.

3. Three words before and three words after the roots of the token.

4. The type/kind of tokens for learning punctuations 4 words before and 4 words after the term.

The ranges of features specified above facilitated the system in learning long and multiword concepts, while considering granularity levels, from the Development (training) dataset.

The results of evaluation of the SVM-based system, using the Test dataset 1, are tabulated in Table 7-7. These results indicate that the SVM-based system achieved high precision rates, generally. However, the corresponding recall rates were generally low. This can be attributed to the fact that the SVM classifier was unable to predict the correct levels of granularity of the annotated concepts in the Test dataset 1.

The SVM-based system outperformed the baseline system by achieving high f-measures for the individual semantic categories, as depicted in Figure 7-5. The exceptions to this are the

semantic categories 'Attribute', 'Findings', 'Observable Entity', 'Qualifier Value', 'Regime/Therapy' and 'Situation'.

Table 7-7: Evaluation of SVM-based system against gold standard Test dataset 1.

Semantic Categories	**True positives**	**Actual concepts in Gold Standard**	**Recall (%)**	**Precision (%)**	**F-measure (%)**
Attribute	213	299	71	89	79
BodyStructure	19	25	73	73	73
Disorder	77	167	46	92	61
Environment	38	74	51	95	67
Findings	89	244	36	76	49
ObservableEntity	55	171	32	60	42
Occupation	40	68	59	89	71
Organism	**Not present**				
Person	173	200	86	98	92
PhysicalObject	9	40	22	60	33
Procedure	73	125	58	65	62
ProductorSubstance	158	222	71	81	76
QualifierValue	348	848	41	75	53
RecordArtifact	3	17	18	50	26
Regime/Therapy	12	96	12	75	21
Situation	16	76	21	100	35
Micro summary	1323	**2672**	49	81	61

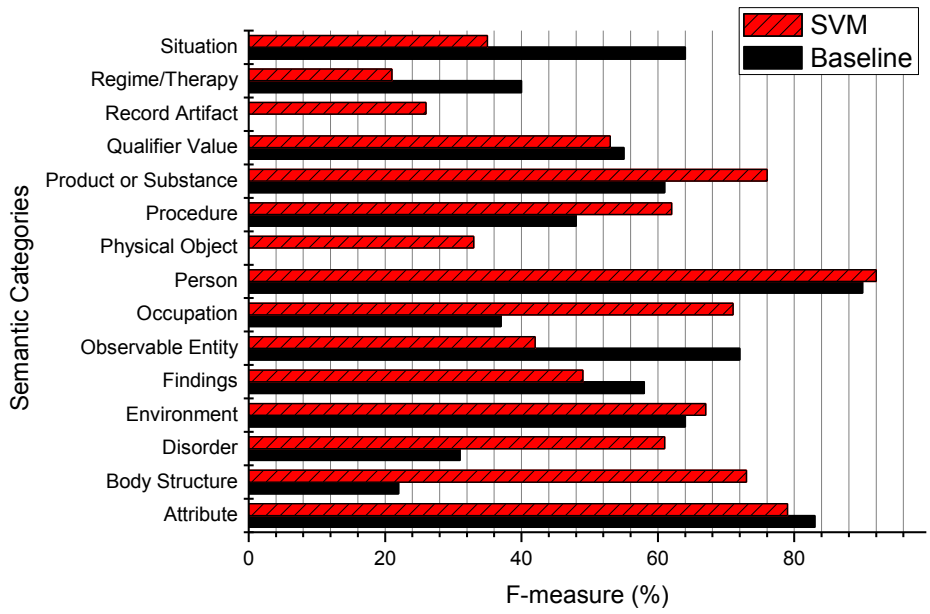


Figure 7-5: Comparison of f-measures of baseline system and SVM-based system on Test dataset 1.

This was due to the lack of training examples for each semantic category and the data imbalance problem that resulted in the biased prediction by the SVM classifier. In addition, the Test dataset 1 was different from the training data (Development dataset) in terms of its content and its format. This might also be considered as a reason for the overall low performance of the SVM-based system.

The Test dataset 2 was similar in nature to the training data. Therefore, the performance of the SVM-based system was expected to be better on the Test dataset 2. However, from the results of evaluation, which are presented in Table 7-8, it is evident that the SVM-based system did not achieve high f-measures for all of the semantic categories.

Table 7-8: Evaluation of SVM-based system against gold standard Test dataset 2.

Semantic categories	**True positives**	**Actual concepts in Gold Standard**	**Recall (%)**	**Precision (%)**	**F-measure (%)**
Attribute	1168	1500	78	84	**81**
BodyStructure	584	1445	40	93	**56**
Disorder	527	1442	37	95	**53**
Environment	333	532	63	97	**76**
Findings	599	2958	20	84	33
ObservableEntity	115	1195	10	85	17
Occupation	68	222	31	80	44
Organism	13	57	23	43	30
Person	533	654	81	99	89
PhysicalObject	295	630	47	86	61
Procedure	903	2300	39	96	56
ProductorSubstance	734	1619	45	92	**61**
QualifierValue	2143	5440	39	82	**53**
RecordArtifact	86	130	66	97	79
Regime/Therapy	31	408	8	78	14
Situation	35	321	11	50	**18**
Micro Summary	8167	**20853**	39	88	54

In comparison with the baseline results, the SVM-based system achieved low f-measure for the semantic categories ‘Attribute’, ‘Disorder’, ‘Findings’, ‘Organism’, ‘Product Or Substance’, ‘Qualifier Value’ and ‘Situation’ in the Test dataset 2. This is also shown in Figure 7-6. The

reasons for this include the small size of the training data and the variation in writing styles that were found in Test dataset 2.

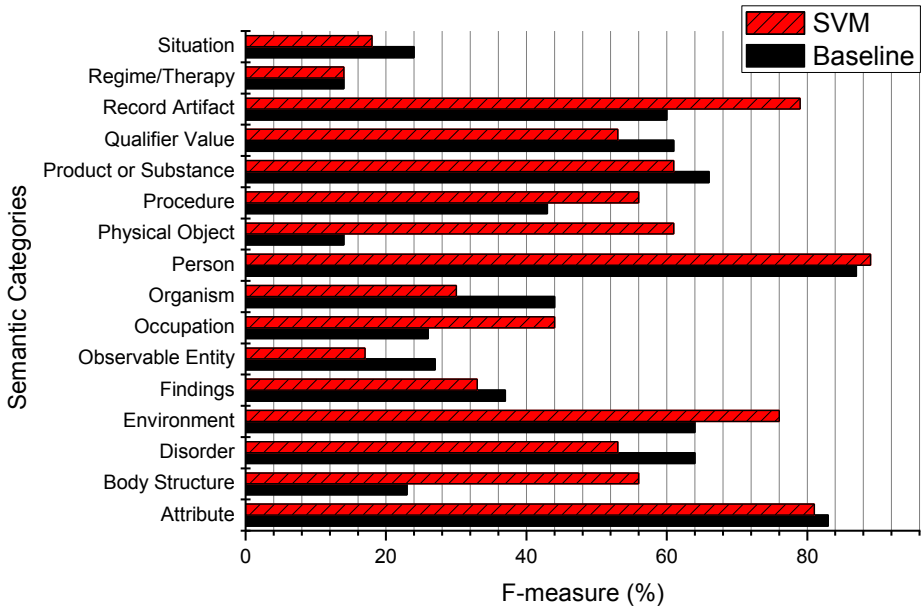


Figure 7-6: Comparison of f-measures of baseline system and SVM-based system on Test dataset 2.

In summary, it can be concluded that the primary contributor in the overall low performance of the SVM-based system was the lack of training examples in the training data (Development dataset). After the evaluation of all approaches on both gold standard test datasets, the performance of each system was compared with the performance of SnoMedTagger and this is discussed in the next section.

7.6 Comparison of rule-based SnoMedTagger with other systems

In the preceding sections of this chapter, the results of evaluation of the baseline system, the SnoMedTagger, the BioPortal web annotator and the SVM-based system are considered. In this section, the performance of the rule-pattern-based SnoMedTagger is compared with the performance of the other three systems. Table 7-9 provides the recall, precision and f-measure, achieved by the various systems, for the individual semantic categories in the Test dataset 1.

The individual comparisons of performance measurements for each system are also presented in Appendix B.

Table 7-9: Comparison of SnoMedTagger with baseline application, BioPortal web annotator and SVM-based system using Test dataset 1.

	Application											
	Baseline			**SnoMed-Tagger**			**BioPortal**			**SVM**		
Semantic categories	Recall (%)	Precision (%)	F-measure (%)	Recall (%)	Precision (%)	F-measure (%)	Recall (%)	Precision (%)	F-measure (%)	Recall (%)	Precision (%)	F-measure (%)
Attribute	95	74	83	93	80	86	25	43	31	71	89	79
BodyStructure	15	40	22	92	65	76	82	30	44	73	73	73
Disorder	19	86	31	75	81	78	64	39	48	46	92	61
Environment	48	96	64	65	87	74	35	86	50	51	95	67
Findings	46	78	58	70	59	64	48	34	40	36	76	49
ObservableEntity	75	69	72	70	70	70	53	26	35	32	60	42
Occupation	26	67	37	87	73	79	89	62	73	59	89	71
Organism	**Not present**											
Person	83	98	90	98	98	98	81	61	70	86	98	92
PhysicalObject	0	0	0	95	75	84	10	13	12	22	60	33
Procedure	34	83	48	86	50	63	41	36	38	58	65	62
ProductorSubstance	64	58	61	86	79	83	88	48	62	71	81	76
QualifierValue	58	52	55	82	65	72	46	36	41	41	75	53
RecordArtifact	0	0	0	59	71	65	25	100	40	18	50	26
Regime/Therapy	28	70	40	84	78	81	21	52	30	12	75	21
Situation	51	86	64	79	79	79	67	72	69	21	100	35
Micro summary	58	65	62	**82**	**71**	**76**	52	40	45	49	**81**	61

It was noted that the overall precision of the SVM-based system was considerably higher compared to the precision of other systems. However, due to the generally low recall rates achieved by this system for the individual semantic categories, the overall f-measure was low. This is because of small number of training examples in the Development dataset (training data). In addition to this, Test dataset 1 was different from training data which also contributed as a reason of low performance of the SVM-based system.

In case of SnoMedTagger, it is instructive to mention here that the Test dataset 1 was completely different from the Development dataset in terms of content and format. Thus, the

generally high recall, precision and f-measure, achieved as a result of evaluation using Test dataset 1, established the general applicability of the SnoMedTagger (Hina, Atwell and Johnson 2013b).

For the four systems, the results of evaluation using the Test dataset 2 are provided in Table 7-10. The individual comparisons of performance measurements for each system are also presented in Appendix C.

Table 7-10: Comparison of SnoMedTagger with baseline application, BioPortal web annotator and SVM-based system using Test dataset 2.

	Application											
	Baseline			**SnoMed-Tagger**			**BioPortal**			**SVM**		
Semantic categories	Recall (%)	Precision (%)	F-measure (%)	Recall (%)	Precision (%)	F-measure (%)	Recall (%)	Precision (%)	F-measure (%)	Recall (%)	Precision (%)	F-measure (%)
Attribute	91	77	83	99	85	91	35	64	45	78	84	81
BodyStructure	13	86	23	81	97	88	62	52	57	40	93	56
Disorder	48	94	64	74	98	84	64	49	56	37	95	53
Environment	50	87	64	79	94	86	59	54	56	63	97	76
Findings	24	80	37	49	84	62	33	27	30	20	84	33
ObservableEntity	17	68	27	31	81	45	22	43	29	10	85	17
Occupation	15	87	26	62	77	69	28	81	42	31	80	44
Organism	35	61	44	44	34	38	75	29	41	23	43	30
Person	77	99	87	90	99	94	77	75	76	81	99	89
PhysicalObject	7	72	14	77	81	79	24	69	35	47	86	61
Procedure	27	96	43	80	96	87	35	49	41	39	96	56
ProductorSubstance	53	86	66	70	98	82	60	36	45	45	92	61
QualifierValue	66	57	61	99	85	91	62	33	43	39	82	53
RecordArtifact	43	100	60	95	87	91	22	94	36	66	97	79
Regime/Therapy	8	64	14	18	88	30	6	35	11	8	78	14
Situation	17	42	24	73	46	56	23	48	31	11	50	18
Micro summary	44	71	55	**77**	**87**	**82**	48	39	43	39	**88**	54

However, Test dataset 2 contained large number of concepts associated with each semantic category; SnoMedTagger was able to achieve high f-measure in comparison with the other three systems. This ensured that the rule-patterns are generally applicable on different datasets.

In terms of the overall comparison on the basis of the micro summaries achieved on both test datasets, presented in Figure 7-7 and Figure 7-8, the following conclusions can be drawn. From overall comparison, it is clear that the SnoMedTagger outperformed the other three systems on both test datasets regardless of the size and nature.

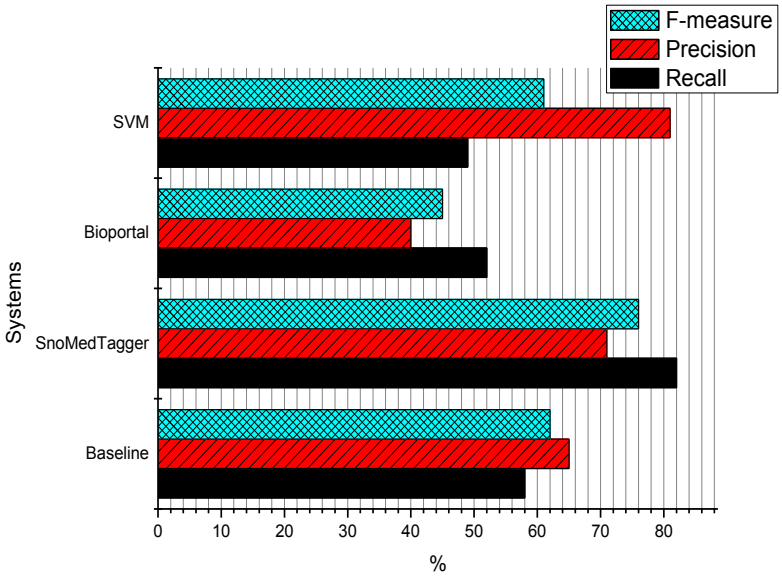


Figure 7-7: Overall performance of various systems achieved for Test dataset 1.

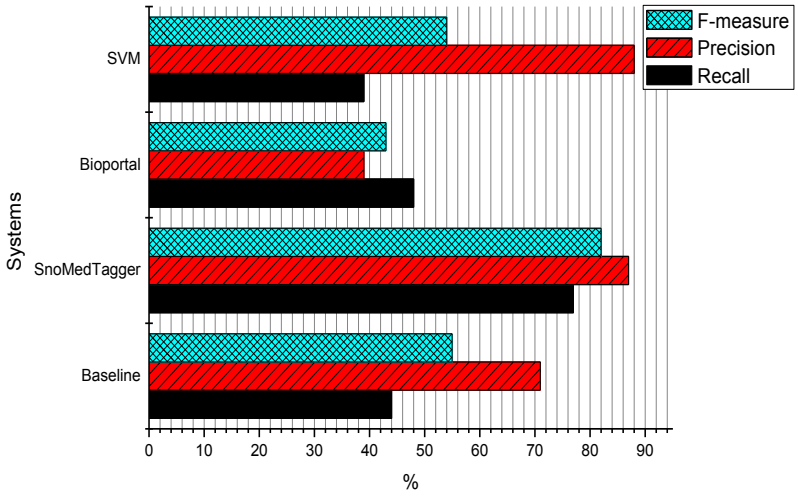


Figure 7-8 Overall performance of various systems achieved for Test dataset 2.

In comparison to the other systems that were considered, the application SVM-based system achieved higher overall precision for the two datasets. However, the recall rates were considerably lower. This can be attributed due to the granularity levels which were not achieved within the given constraint that is the small training data. This resulted in less training for each individual semantic category. Therefore, it can be concluded that large and balance training data can contribute to the general applicability of machine learning approach on different datasets.

As expected, the overall performance of the ontology-based Bioportal web annotator was inferior in comparison to the performance of the other systems. This is valid for both the Test dataset 1 and Test dataset 2 and it can be considered as a clear indication of the fact that concepts in the SNOMED CT clinical controlled vocabulary are insufficient to identify semantic information in medical narratives. Although, the SNOMED CT clinical vocabulary cannot be directly incorporated within medical narratives, it still served as a useful resource to recognise the gap between controlled vocabularies and medical narratives.

7.7 Validation of the output of the SnoMedTagger

It was considered to be important to validate the output of the SnoMedTagger. Output here refers to the semantic information identified by the SnoMedTagger and classified into various semantic categories. For validation, two general practitioners[10] (who were not the original annotators) checked the following aspects in the output of the SnoMedTagger.

- Classification of a concept in the correct semantic category.
- Possibility of a concept belonging to more than one semantic category.
- Identification of the complete boundaries of a concept.

The rule-patterns and the dictionaries were updated in the light of the feedback given by the general practitioners. Some of the more important points that were raised and the measures taken to deal with the same are now presented.

[10] 1. Dr. Marc Jamoulle, Family doctor, Health data management specialist.
2. Dr. Richard Gwent Jones, Consultant, Leeds Teaching Hospitals NHS Trust.

Few multiword concepts, such as ‘Chronic RBC’, were incorrectly identified by the SnoMedTagger as ‘Body Structure’. At the same time, ‘RBC’ was correctly identified as ‘Body Structure’. To resolve this problem, the concept value ‘chronic’ was restricted for the rule-pattern that identified ‘chronic RBC’ as ‘Body Structure’. This was done because the rule-pattern “({QualifierValue}{sp}{BodyStructure})” is generic, therefore, it can not be removed from the application. Furthermore, removing ‘chronic’ from the dictionary is likely to affect the general applicability of other rule-patterns.

Classification of concepts such as ‘no lymphadenopathy’ and ‘no hepatosplenomegaly’ in the semantic category ‘Disorder’ was regarded as incorrect. This is because both the general practitioners were of the opinion that ‘no disorder’ is not a disorder. The rule-pattern that identified such concepts is “({QualifierValue}{sp}{Disorder})” where “no” is the “Qualifier Value” and “lymphadenopathy” or “hepatosplenomegaly” are the “Disorder”. This rule-pattern is generic; therefore, instead of removing the rule-pattern, concept value ‘no’ was restricted for this specific pattern.

Furthermore, on the basis of personal experience of writing consultations, the general practitioners (GPs) suggested adding some of the concepts that are commonly used in medical narratives. Examples include ‘internal’, ‘external’, ‘lateral’, ‘lower’, ‘upper’, etc. Since these concepts were already present in the dictionary of ‘Qualifier Value’ and GPs were unaware of the concepts present in dictionaries, no further action was required in this case.

The GPs suggested that concepts such as Aerosol, capsule, suspension, graft, etc., which were classified into the semantic category ‘Product Or Substance’, also belong to the semantic category ‘PhysicalObject’. Therefore, after getting consent of the reviewers annotators (domain experts involved in annotation of gold standard), such concepts were added in the dictionary of ‘Physical Object’.

7.8 Summary

This chapter explained the complete evaluation and validation carried out for the SnoMedTagger. SnoMedTagger was first evaluated using two different gold standard test

datasets (Test dataset 1, Test dataset 2). Then, the performance of the baseline system, the BioPortal web annotator and the SVM-based system was compared with the performance of the rule-based SnoMedTagger.

SnoMedTagger not only improved the f-measures up to 14%-27% against the baseline system using gold standard test datasets but also outperformed the ontology-based BioPortal web annotator and the SVM-based machine learning system. Other than the evaluation of SnoMedTagger against different systems, the output concepts were also validated by two domain experts and on the basis of their feedback, dictionaries and rule-patterns were updated. Although, SnoMedTagger was evaluated and validated properly, still there are some limitations which can be tackled as part of future work, explained in the next chapter.

Chapter 8. Conclusions

The primary aim of this research was achieved by developing SnoMedTagger, which is a generic semantic tagger for medical narratives. The development process of the SnoMedTagger presented a number of challenges. The means devised to tackle these challenges are reported as secondary contributions of this research. This chapter summarises the outcomes of the claimed contributions. In addition, the limitations of this work are stated so is the potential future work.

8.1 Summary of the results

Primary Contribution:

The main contribution of this research is the development of SnoMedTagger – SNOMED CT Medical Tagger. The SnoMedTagger was developed using SNOMED CT, which is the international healthcare clinical terminology. To the best of our knowledge, SnoMedTagger is the first semantic tagger which can be used by researchers to identify semantic information in medical narratives and classify this information into globally known semantic categories that are derived from SNOMED CT. It has been shown in this work that the SnoMedTagger is able to identify and classify individual concepts, paraphrases of concepts, abbreviations of concepts and complex multiword concepts. Details of the development of the SnoMedTagger have been published, so are the results of its performance testing (Hina, Atwell and Johnson 2013b). The SnoMedTagger is available from http://www.comp.leeds.ac.uk/scsh/SnoMedTagger.html.

Secondary Contributions:

The challenges tackled during the development of SnoMedTagger were categorised as secondary contributions. This was done because there were no standard methods existed to tackle these challenges. The results of these challenges are summarised as follows.

- **Methodological evaluation and validation of SnoMedTagger** - To prove the significance and applicability of main contribution, a comprehensive methodological evaluation and validation was carried out in this research which was not previously done for some state-of-

the-art systems such as MetaMap (Aronson and Lang 2010). The evaluation of SnoMedTagger was done on the basis of two different gold standard test datasets that were annotated by two domain experts. On Test dataset 1, SnoMedTagger performed reasonably well and achieved an overall recall of 82%, 71% precision and 76% of f-measure; while on Test dataset 2, it scored 77% recall, a high precision of 87% and 82% on f-measure.

These performance measurements demonstrated the applicability of SnoMedTagger on different datasets. In order to compare the rule-based approach of SnoMedTagger, the results were also compared against other approaches (baseline system - SNOMED CT dictionary application, an ontology-based BioPortal web annotator and an SVM-based machine learning system). SnoMedTagger outperformed all three systems with an improvement in accuracy of 14% against the baseline system, 31% against the existing ontology-based BioPortal web annotator and 15% against the SVM-based machine learning system on Test dataset 1 (Hina, Atwell and Johnson 2013b; Hina, Atwell and Johnson 2013a). On Test dataset 2, SnoMedTagger achieved better f-measure with an improvement of 27% against the baseline system, 39% against the BioPortal web annotator and 28% against the SVM-based system.

In the case of individual performance measurements (recall, precision and f- measure), the SVM-based machine learning system achieved high precision on both test datasets but very low recall because of different levels of granularity in the identification of multiword concepts. This low recall rate decreased the overall f-measure of this method as compared to overall f-measure achieved by SnoMedTagger. In the case of the machine learning method, it was difficult to achieve general applicability because it can only perform better in the case of similar data (training and test). On the other hand, the ontology-based BioPortal web annotator overall achieved low scores on both datasets because of the inappropriateness of controlled vocabularies.

After the evaluation against gold standard test datasets and systems using different approaches, the output concepts from both datasets were validated by two different domain

experts. On the basis of feedback obtained from said domain experts, rules were re-analysed and dictionaries were updated.

- **Anonymization module** – Besides the Development dataset and the Test dataset 2, which were selected from the i2b2 challenge corpus, another test dataset was required to test the performance of SnoMedTagger on different datasets. Therefore, a test dataset, which is referred to as Test dataset 1 in this work, was extracted from 'SystmOne'. The Test dataset 1 was representative of real data and contained fictional information of patients associated with four PHI categories. Therefore, anonymization of the Test dataset 1 was required.

  The existing anonymization systems were developed for corpora of different nature and for different PHI categories. Thus, these systems were considered inappropriate for anonymization of the Test dataset 1, which contained a mixture of natural language and clinical codes. For this reason, an anonymization module was developed for the Test dataset 1. Evaluation against the gold standard 'Evaluation set' showed that the overall f-measure achieved by the developed anonymization module was 24% higher compared to that achieved by the baseline system (Hina et al. 2013). This anonymization module was also contributed in a project that aimed to make real-data available for researchers within cloud based Virtual Research Environment – VRE (Smith et al. 2013).

- **Annotation guidelines for medical narratives -** To develop a gold standard corpus for the development and the evaluation of SnoMedTagger, general annotation guidelines were produced on the basis of language issues (e.g., paraphrases of concepts, abbreviations of concepts, and complex multiword concepts) that were identified by the baseline system (Hina, Atwell and Johnson 2011). For the validation of annotation guidelines, a non-domain user annotated the Development dataset. This annotated dataset was then reviewed by a domain expert who agreed on more than 90% of the annotations. After this validation, two domain experts followed these guidelines and used the semi-automatic approach to annotate the Development dataset and the Test dataset 1. However, the Test dataset 2 was annotated using the manual approach, as explained in Chapter 5. Following the aforementioned procedure, high inter-annotator agreement scores ranging between 86% - 95.25% were

achieved. This indicated that the annotation guidelines can be reliably applied on different datasets.

8.2 Limitations and suggestions for future work

The SnoMedTagger can be employed for extraction of semantic information from medical narratives. However, there are limitations, which are listed as follows, so are the suggested actions to overcome these limitations.

- The annotation guidelines for the annotation of gold standard dataset were used by domain experts and also by non-domain users. However, the inter-annotator agreement was not calculated in the case of non-domain users. Thus, in order to establish the utility of annotation guidelines for non-domain users, it is suggested to compare the inter-annotator agreement scores for non-domain users with the inter-annotator agreement scores for domain experts.
- To achieve general applicability of the SnoMedTagger, all of the annotated concepts in the gold standard and most of the concepts in the dictionaries were refined (Section 6.1.2). The long, multiword concepts in the dictionary that could not be refined due to time constraints were left in their original form. Thus, a proposed direction for further work is to analyse and refine such concepts followed by reformulation of the SnoMedTagger to achieve more detailed levels of granularity.
- In the evaluation of SnoMedTagger on Test dataset 2, it was found that the application did not achieve impressive f-measures for semantic categories ‘Findings’, ‘Observable Entity’, ‘Regime/Therapy’ and ‘Situation’. This leads to another potential future work which will involve the evaluation of SnoMedTagger on specific test cases. Such test cases, preferably designed by domain experts, should contain multiword concepts selected from different datasets. The evaluation of the performance of SnoMedTagger will provide a basis for improving the generic rule-patterns for the above mentioned semantic categories.

- Since the performance of SVM based system (Section 7.5) was restricted due to small size of training set. Another interesting future work could involve using whole of the data used in this research as training set and obtaining another dataset as test set.
- It is important to note that SNOMED CT is updated annually. As a result, the dictionaries used in the SnoMedTagger may need to be updated in case of addition of new concepts in the SNOMED CT.
- On the basis of the limitations listed below, the evaluation of the anonymization module on different corpora and its modification (if required) to make it generally applicable are suggested as potential future work.
    - The module was developed for a specific corpus (Test dataset 1) that contained a mixture of medical narratives and clinical codes. In addition, the four PHI categories in this module are specific to the corpus that was used in this work.
    - Due to resource constraints, the performance of the rule-based anonymization module was not evaluated against any other corpus. For the same reason, the performance of the rule-based approach was not compared with the performance of other approaches such as machine learning approach.

The potential applications of SnoMedTagger are also considered as part of future work. This involves using SnoMedTagger for developing question-answering systems such as ‘finding cause of death in verbal autopsies’. For this research question, the SnoMedTagger was explored for the extraction of features to be used in a machine learning system (Danso et al. 2013). These features are relevant semantic categories which will contribute in training a classifier for finding cause of death in verbal autopsies. Another future application for the above mentioned research question is an extended rule-based system which will use the semantic categories ‘Findings’ and ‘Disorder’. Using the SnoMedTagger, these semantic categories will be extracted as potential features to identify ‘cause of death’ in verbal autopsies. The results will then be evaluated against machine learning system.

SnoMedTagger can also be used in finding relationships between relevant semantic information in medical narratives. For instance, findings the relationship between ‘Disorder’ and ‘Product or Substance’ for answering research question such as ‘Which medication is prescribed for which disease?’ or ‘Finding diagnosis and treatment information in patient’s consultation notes’, etc.

A researcher at Kyung Hee University (South Korea) has contacted me to use our semantic tagger in her research on interoperability of concepts in discharge summaries and SNOMED CT healthcare terminology. In particular, her research is to investigate the ability of a system to exchange information between SNOMED CT healthcare terminology codes and the text written in discharge summaries. This will also lead to a future research application that will extract semantic information from medical narratives using SnoMedTagger and then use it to codify relevant concepts in SNOMED CT health care terminology.

In addition, a medical doctor also wants to use SnoMedTagger for knowledge extraction from health records that are based on SNOMED CT. SnoMedTagger was implemented as a GATE application and relied on using other GATE components. Therefore, contributing it to the GATE open source tool is one way of making it available for the research community.

An enormous amount of patient’s data exists in textual reports. This data needs to be processed and encoded in timely manner. This will require domain experts to manually analyse and encode important information in the text. This approach is time consuming and impractical. The use of SnoMedTagger will also be helpful in reducing human effort to analyse important information that resides in large volumes of patient records. However, the reliability of this analysis needs prior check on the selection of best performing semantic categories.

On the basis of evaluation and practical applications discussed in this thesis, it can be concluded that the SnoMedTagger can be used by researchers working on research questions that involve medical narratives.

Appendices

Appendix A : SNOMED CT fact sheet

SNOMED CT Semantic Categories	Description	Examples
Attribute	Attribute is the sub-class of top-level concept class 'linkage concept'. The concepts in this category are used to construct relationships between SNOMED CT concepts which can then be used to define the logical meaning of a concept.	Associated with, after, causing, date, due to, during, etc.
Body Structure	Concepts in this category include normal and abnormal anatomical structures. Normal anatomical structures specify the body site involved by a disease or procedure.	Zone of lung, heart tissue, ear structure, ear hair, entire heart, etc.
Disorder	Disorder is the sub-class of top-level concept category 'clinical findings'. Concepts under this category are descendants of 'disease' and refer to abnormal clinical states.	Tuberculosis, burn shock, busitis of hand, buruli ulcer.
Environment	This semantic category contains all types of environments and locations.	Home, hospital, warehouse, yard, zoo, I.C.U., etc.
Findings	Concepts which are results of clinical observations or examinations. These include normal and abnormal clinical states.	Able to run, absence of toe, anxiety, death, etc.
Observable Entity	This top-level semantic category represents question or procedure which can produce an answer or a result. These entities can also be used as an element where a value can be assigned. For instance, | Left ventricular end-diastolic pressure (observable entity) | could be interpreted as the question, "What is the left ventricular end diastolic pressure?" or "What is the measured left ventricular end-diastolic pressure?" Observables are entities that could be used to code elements on a checklist or any element where a value can be assigned. | Color of nail (observable entity) | is an observable. | Gray nails (finding) | is a finding. One use for | Observable entity | in a clinical record is to code headers on a template. For example, | Gender (observable entity) | could be used to code a section of a template titled "Gender" where the user would choose "male" or "female". "Female gender" would then constitute a finding.	'colour of nail', 'age', 'gender', 'length of ulna', 'blood pressure', etc.
Occupation	It is a sub-class of the top-level concept class 'social context' and contains all concepts which are occupations.	'doctor', 'general practitioner', 'nurse', 'clerk',

SNOMED CT Semantic Categories	Description	Examples
		'manager', 'actor', etc.
Organism	Concepts in this category include organisms of significance in human and animal medicine or in modelling the causes of diseases.	'algae', 'alnus', 'amoeba', 'black fly', 'cryptocotyle', etc.
Person	It is another sub-class of the top-level concept category 'social context' and contains concepts which can be referred to as a person.	'employer', 'patient', 'baby', 'father', etc.
Physical Object	Concepts in this category include natural or man-made objects or objects used to model the concepts in the 'procedure' category.	'book', 'needle', 'boiler', 'cloth', etc.
Procedure	Concepts in this category include activities performed in the provision of health care.	'radiography', 'measles vaccination', 'operation on the ear', 'optimal surgery', etc.
Product Or Substance	For the present study, two top-level concept categories 'pharmaceutical/biological product' and 'substance' were combined to form this semantic type. This was done on the basis of observation that these two semantic types were interchangeably used frequently in the medical narratives. However, in SNOMED CT the concept category 'pharmaceutical/biological product' contains drug products and 'substance' contains chemical constituents of drug products (in the 'pharmaceutical/biological product' category), food and chemical allergens, adverse reactions and toxicity information	'vancomycin' (Product), 'VAL syrup', 'topical from Zinc' (Product), sodium citrate (substance), etc.
Qualifier Value	The | Qualifier value | hierarchy contains some of the concepts used as values for SNOMED CT attributes that are not contained elsewhere in SNOMED CT. Such a code may be used as the value of an attribute in a defining Relationship in pre-coordinated definitions, and/or as the value of an attribute in a qualifier in a post-coordinated expression. However, the values for attributes are not limited to this hierarchy and are also found in hierarchies other than | Qualifier value |. For example, the value for the attribute | LATERALITY | in the concept shown below is taken from the | Qualifier value | hierarchy: • | Left kidney structure | | LATERALITY | | Left |. However, the value for the attribute | FINDING SITE | in the concept shown below is taken from the | Body	'left', 'right', 'first', 'upper', 'unit of rate', 'simple', etc.

SNOMED CT Semantic Categories	Description	Examples
	Structure | hierarchy, not the | Qualifier value | hierarchy. • | Pneumonia | | FINDING SITE | | Lung structure |.	
Record Artifact	Concepts in this category are entities created by a 'person' to provide information on events or records.	'death summary', 'discharge summary', 'summary report', 'radiology report', etc.
Regime/Therapy	It is a sub-class of top-level category 'procedure' and includes concepts focal in the 'procedure'.	'art therapy', 'cold therapy, 'ear care', dying care', etc.
Situation	Concepts in the | Procedure | and |Clinical finding| hierarchies (given the appropriate record structure) can be used in a clinical record to represent: • Conditions and procedures that have not yet occurred (e.g. | Endoscopy arranged (situation) |); • Conditions and procedures that refer to someone other than the patient (e.g. | Family history: Diabetes mellitus (situation) |, | Discussed with next of kin (situation) |); • Conditions and procedures that have occurred at some time prior to the time of the current entry in the record (e.g. | History of - aortic aneurysm (situation) |, |History of - splenectomy (situation)|). In each of these examples, clinical context is specified. The second example, in which someone other than the patient is the focus of the concept, could be represented in an application or record structure by combining a header term Family history with the value Diabetes. The specific context (in this case, family history) would be represented using the record structure. In this case, the pre-coordinated context-dependent concept | Family history: Diabetes mellitus (situation) | would not be used because the information model has already captured the family history aspect of the diabetes. Concepts in the | Procedure | and |Clinical finding| hierarchy have a default context of the following: • The procedure has actually occurred(versus being planned or cancelled) or the finding is actually present (versus being ruled out, or considered); • The procedure or finding being recorded refers to the patient of record (versus, for example, a	'history of anemia', 'family history', 'no nausea', etc.

SNOMED CT Semantic Categories	Description	Examples
	family member); • The procedure or finding is occurring now or at a specified time (versus some time in the past). In addition to using the record structure to represent context, there is sometimes a need to override these defaults and specify a particular context using the formal logic of the terminology. For that reason, SNOMED CT has developed a context model to allow users and/or implementers to specify context using the terminology, without depending on a particular record structure. The | Situation with explicit context | hierarchy and various attributes assigned to concepts in this hierarchy accomplish this.	

Appendix B: Performance of various systems with respect to each semantic category in Test dataset 1

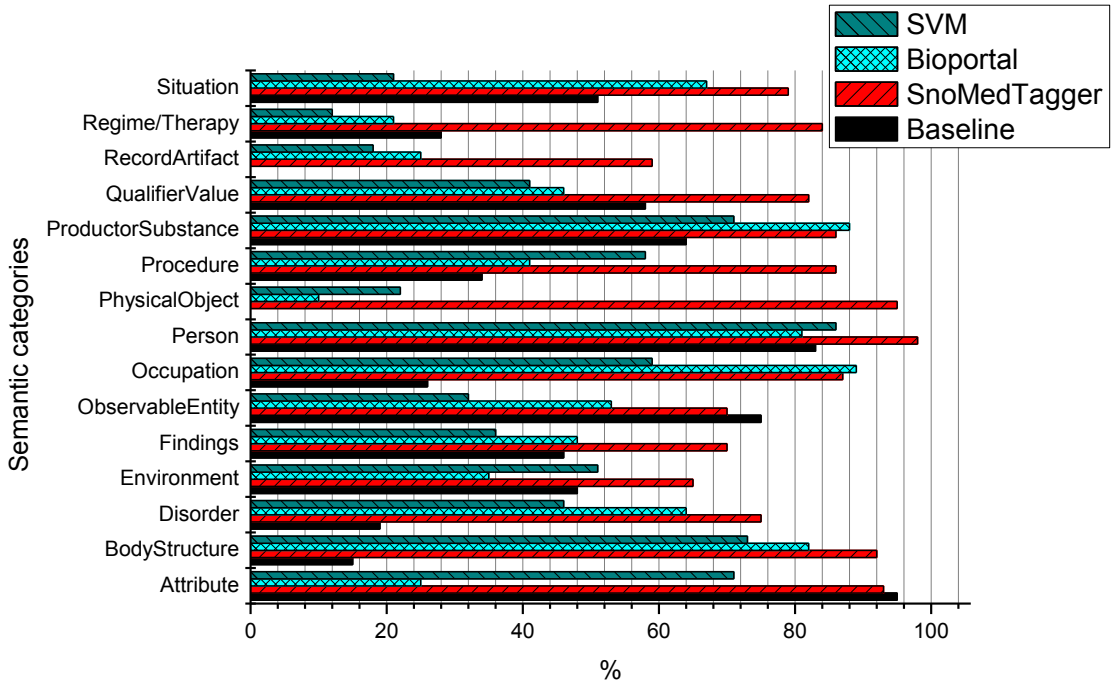


Figure B - 1: Overall recalls (%) achieved by various systems on Test dataset 1.

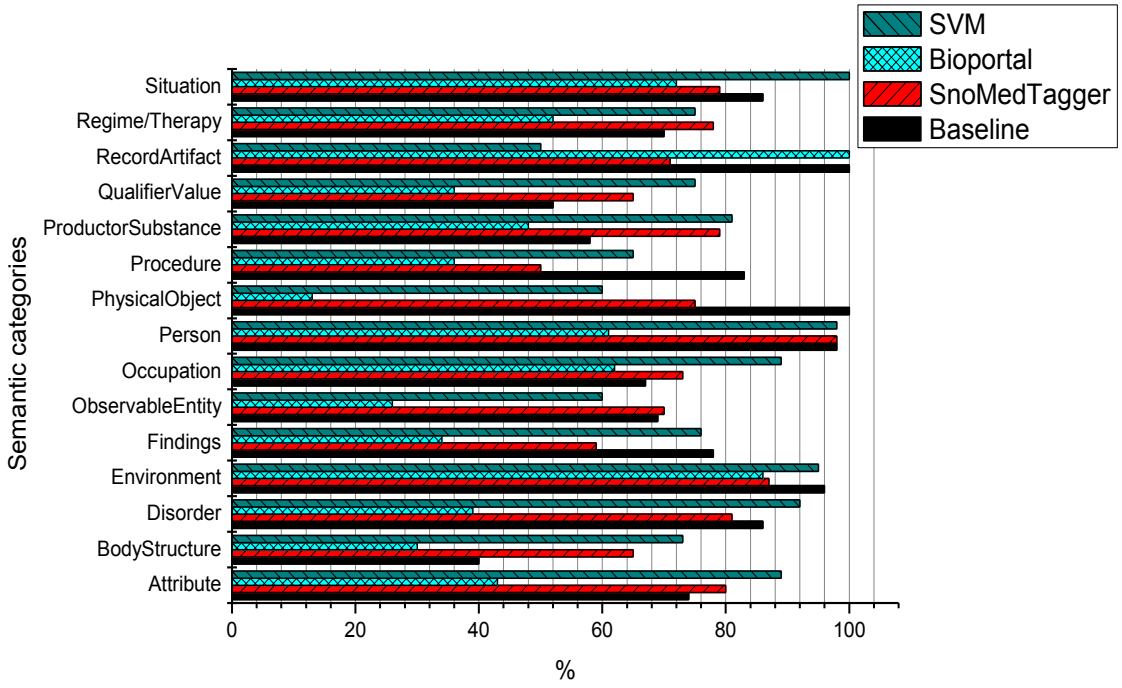


Figure B - 2: Overall precisions (%) achieved by various systems on Test dataset 1.

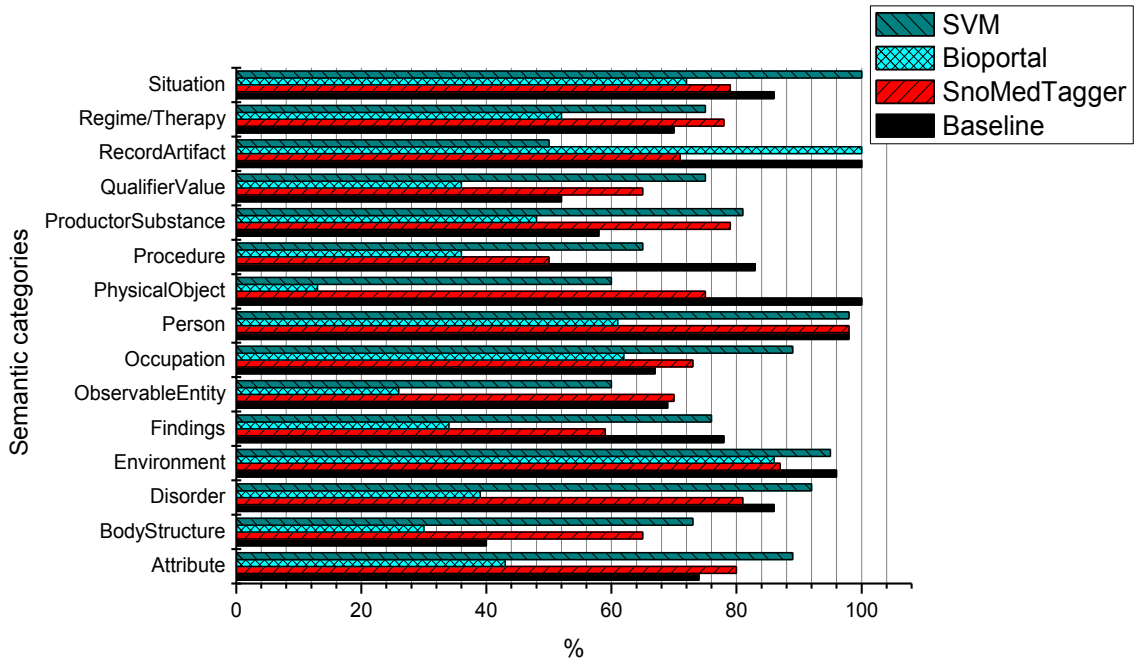


Figure B - 3: Overall f-measures (%) achieved by various systems on Test dataset 1.

Appendix C: Performance of various systems with respect to each semantic category in Test dataset 2

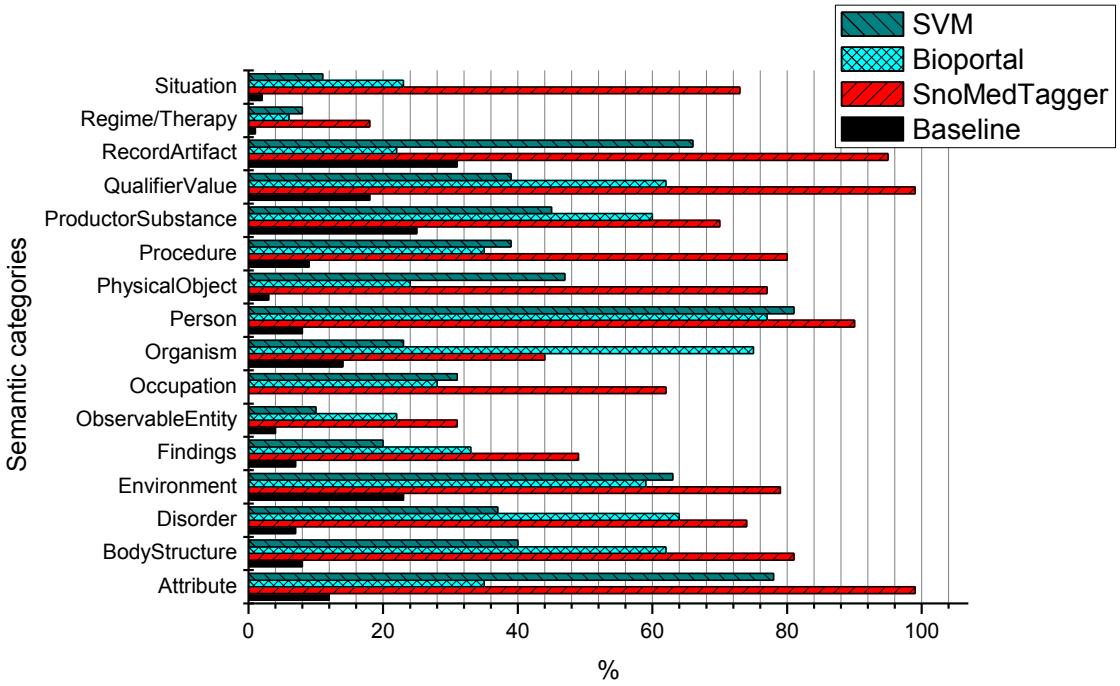


Figure C - 1: Overall recalls (%) achieved by various systems on Test dataset 2.

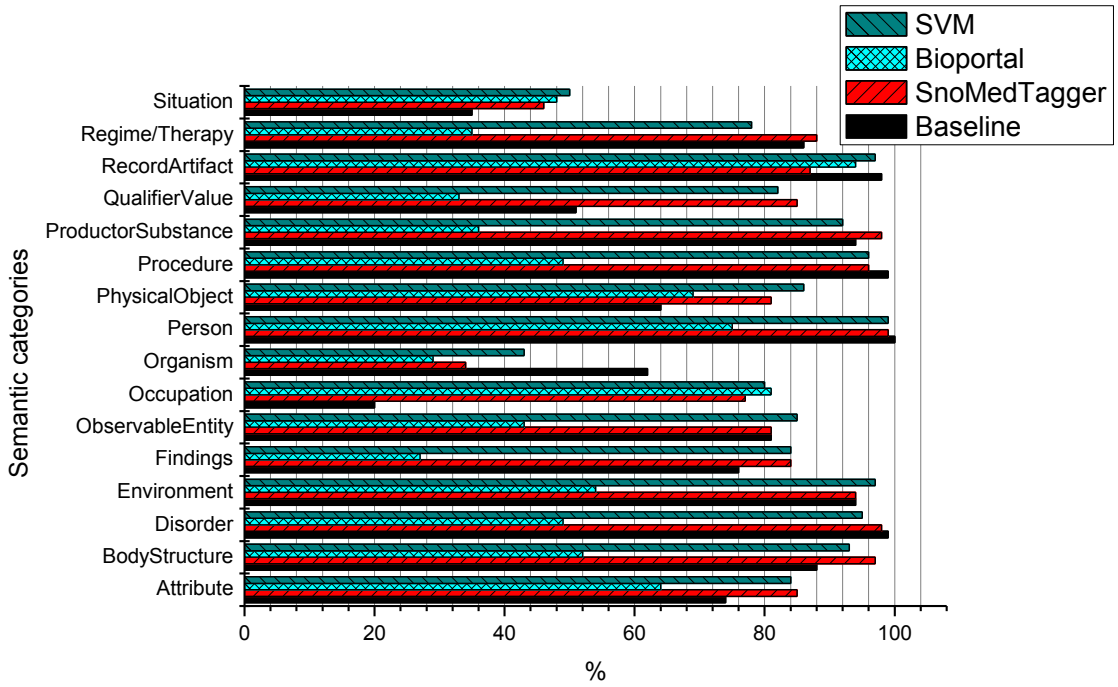


Figure C - 2: Overall precisions (%) achieved by various systems on Test dataset 2.

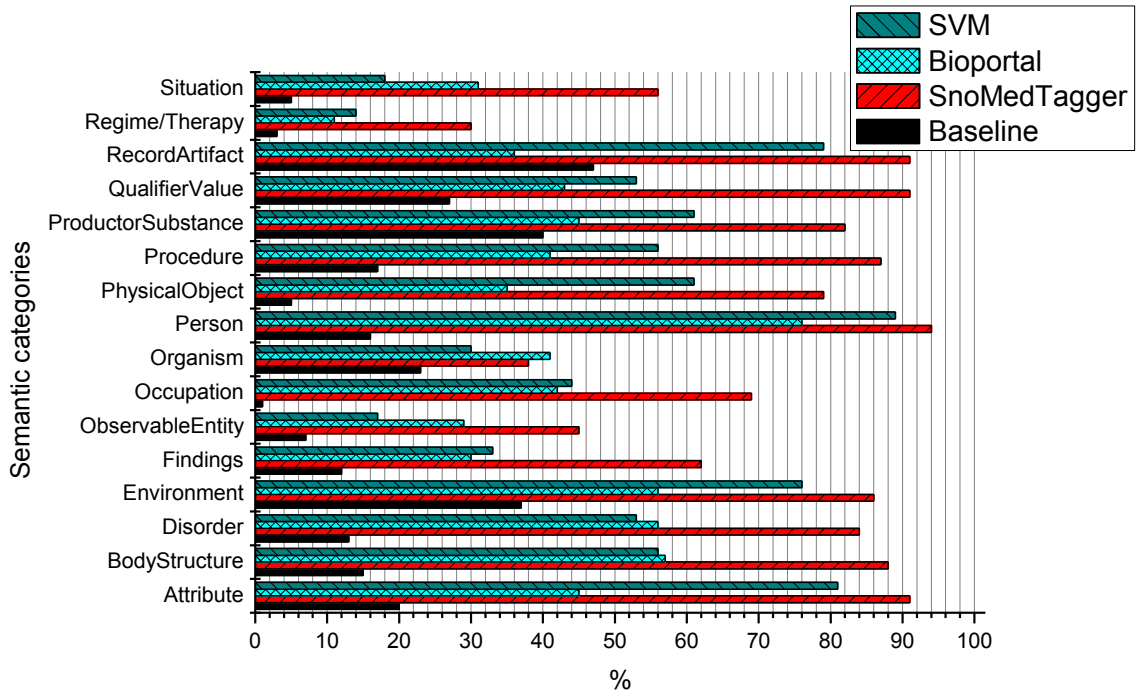


Figure C - 3: Overall f-measure (%) achieved by various systems on Test dataset 2.

References

Abacha, A. B. and Zweigenbaum, P. 2011. Medical Entity Recognition: A Comparison of Semantic and Statistical Methods. *In: Workshop on Biomedical Natural Language Processing, ACL-HLT 2011*, June 23-24, Portland, Oregon, USA. Association for Computaional Linguistics, pp.56-54.

Albright, D., Lanfranchi, A., Fredriksen, A., Styler, W. F., Warner, C., Hwang, J. D., Choi, J. D., Dligach, D., Nielsen, R. D., Martin, J., Ward, W., Palmer, M. and Savova, G. K. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*.

Ananiadou, S., Sullivan, D., Black, W., Levow, G.-A., Gillespie, J. J., Mao, C., Pyysalo, S., Kolluru, B., Tsujii, J. and Sobral, B. 2011. Named Entity Recognition for Bacterial Type IV Secretion Systems. *PLoS ONE*, **6**(3), pe14780.

Ao, H. and Takagi, T. 2005. Alice: An Algorithm to Extract Abbreviations from MEDLINE. *J Am Med Inform Assoc*, **12**, pp.576 - 586.

Appelt, D. E. and Onyshkevych, B. 1998. The common pattern specification language. *In: Proceedings of a workshop on held at Baltimore, Maryland, October 13-15, Baltimore, Maryland*. Association for Computational Linguistics, pp.23-30.

Aramaki, E., Imai, T., Miyo, K. and Ohe, K. 2006. Automatic Deidentification by using Sentence Features and Label Consistency. *In: i2b2 Workshop on challenges in Natural Language Processing for Clinical data: Deidentification and Smoking Challenge*.

Aronson, A. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings AMIA Symposium*, pp.17 - 21.

Aronson, A. R. and Lang, F. M. 2010. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc*, **17**(3), pp.229-36.

Atwell, E. 1983. Constituent-Likelihood Grammar. *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*, **7**, pp.34-67.

Atwell, E. 2008. *Development of tag sets for part-of-speech tagging* [online]. Mouton de Gruyter. Available from: http://www.degruyter.com/view/books/9783110211429/9783110211429.4.501/9783110211429.4.501.xml.

Atwell, E., Demetriou, G., Hughes, J., Schiffrin, A., Souter, C. and Wilcock, S. 2000. A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*, **24**, pp.7-23.

Bashyam, V., Divita, G., Bennett, D. B., Browne, A. C. and Taira, R. K. 2007. A normalized lexical lookup approach to identifying UMLS concepts in free text. *Stud Health Technol Inform,* **129**(Pt 1), pp.545-9.

Bashyam, V. and Taira, R. K. 2005. Indexing anatomical phrases in neuro-radiology reports to the UMLS 2005AA. *AMIA : Annual Symposium proceedings / AMIA Symposium*, pp.26-30.

Baud, R. H., Rassinoux, A. M. and Scherrer, J. R. 1992. Natural language processing and semantical representation of medical texts. *Methods Inf Med,* **31**(2), pp.117-25.

Baud, R. H., Rassinoux, A. M., Wagner, J. C., Lovis, C., Juge, C., Alpay, L. L., Michel, P. A., Degoulet, P. and Scherrer, J. R. 1995. Representing clinical narratives using conceptual graphs. *Methods Inf Med,* **34**(1-2), pp.176-86.

Beckwith, B. 2006. Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC Med Inform Decis Mak*, p12.

Berman, J. 2003. Concept-match medical data scrubbing. How pathology text can be used in research. *Arch Pathol Lab Med*, pp.680 - 6.

Bikel, D. M., Schwartz, R. and Weischedel, R. M. 1999. An Algorithm that Learns What's in a Name. *Mach. Learn.,* **34**(1-3), pp.211-231.

Boufaden, N. 2003. An ontology-based semantic tagger for IE system. *In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 2, Sapporo, Japan*. Association for Computational Linguistics, pp.7-14.

Brierley, C. and Atwell, E. 2010. ProPOSEC: A Prosody and PoS Annotated Spoken English Corpus. *In: LREC'2010 Language Resources and Evaluation Conference*.

Brierley, C., Atwell, E., Rowland, C. and Anderson, J. 2013. Semantic Pathways: a Novel Visualisation of Varieties of English. *ICAME Journal of the International Computer Archive of Modern English,* **37**, pp.5-36.

Brill, E. 1992. A simple rule-based part of speech tagger. *In: Proceedings of the workshop on Speech and Natural Language, Harriman, New York*. Association for Computational Linguistics, pp.112-116.

Brill, E. 1994. Some Advances in transformation-based part of speech tagging. *In: Proceedings of the National Conference on Artificial Intelligence*: AAAI Press, pp.722-727.

Brill, E. 1995. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Comput. Linguist.,* **21**(4), pp.543-565.

Bruijn, B. D., Cherry, C., Kiritchenko, S., Martin, J. and Zhu, X. 2010. NRC at i2b2: one challenge, three practical tasks, nine statistical systems, hundreds of clinical records, millions of useful features. *In: Proceedings of the 2010 i2b2/VA Workshop on Challenges in Natural Language Processing for Clinical Data*, Boston, USA.

Carletta, J. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics,* **22**(2), pp.249-254.

Chapman, W., Bridewell, W., Hanbury, P., Cooper, G. and Buchanan, B. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform,* **34**(5), pp.301 - 10.

Christensen, L. M., Harkema, H., Haug, P. J., Irwin, J. Y. and Chapman, W. W. 2009. ONYX: a system for the semantic analysis of clinical text. *In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, Boulder, Colorado*. Association for Computational Linguistics, pp.19-27.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement,* **20**(1), pp.37-46.

Coiera, E. 2003. Healthcare terminologies and classification systems. *Guide to Health Informatics* 2nd ed. London: Hodder Arnold pp.210-212.

Crossfield, S. and Clamp, S. 2013a. Electronic Health Records Research in a Health Sector Environment with Multiple Provider Types. *In: HEALTHINF 2013 Proceedings of the International Conference on Health Informatics*.

Crossfield, S. S. R. and Clamp, S. E. 2013b. Centralised Electronic Health Records Research across Health Organisation Types. (In publication) *In: BIOSTEC 2013,* Berlin. Communications in Computer and Information Science: Lecture Notes. Springer-Verlag.

Cunningham, H., Mayard, D., Bontcheva, K. and Tablan, V. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *In: 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02),* July 2002, Phildelphia.

Cunningham, H., Mayard, D. and Tablan, V. D. O. C. SCIENCE. 2000. *JAPE: a JAVA Annotation Patterns Engine* (CS--00--10). Sheffield: University of Sheffield.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y. and Peters, W. 2011. *Text Processing with GATE (Version 6).*

Danso, S., Atwell, E., Johnson, O., Ten Asbroek, A., Soromekun, S., Edmond, K., Hurt, C., Hurt, L., Zandoh, C., Tawiah, C., Fenty, J., Etego, S., Agyei, S. and Kirkwood, B. 2013. A Semantically Annotated Verbal Autopsy Corpus for Automatic Analysis of Cause of Death. *ICAME Journal of the International Computer Archive of Modern English,* **37**.

Demetriou, G. and Atwell, E. 2001. A domain-independent semantic tagger for the study of meaning associations in English text. *In: Proceedings of the Fourth International Workshop on Computational Semantics (IWCS-4),* pp.67-80.

Demner-Fushman, D., Chapman, W. W. and Mcdonald, C. J. 2009. What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics,* **42**(5), pp.760-772.

Divita, G., Tse, T. and Roth, L. 2004. Failure analysis of MetaMap Transfer (MMTx). *Medinfo,* **11**(Pt 2), pp.763 - 7.

Dukes, K. and Atwell, E. 2012. LAMP: A Multimodal Web Platform for Collaborative Linguistic Analysis. *In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).European Language Resources Association (ELRA)*.

Dukes, K., Atwell, E. and Habash, N. 2013. Supervised collaboration for syntactic annotation of Quranic Arabic. *Language Resources and Evaluation,* **47**(1), pp.33-62.

F. Baader and Nutt., W. 2002. Basic Description Logics. *In:* F. BAADER, D. CALVANESE, D.L. MCGUINNESS, D. NARDI and P. F. PATEL-SCHNEIDER, eds. *Description Logic Handbook.* Cambridge University Press, pp.47-100.

Feng, D., Burns, G., Zhu, J. and Hovy, E. 2008. Towards Automated Semantic Analysis on Biomedical Research Articles *In: International Joint Conference on Natural Language Processing*.

Friedlin, F. and Mcdonald, C. 2008. A software tool for removing patient identifying information from clinical documents. *J Am Med Inform Assoc,* **15**(5), pp.601 - 10.

Friedman, C. 2005. Semantic Text Parsing for Patient Records. *In:* H. CHEN, S. FULLER, C. FRIEDMAN and W. HERSH, eds. *Medical Informatics.* Springer US, pp.423-448.

Friedman, C., Liu, H., Shagina, L., Johnson, S. and Hripcsak, G. 2001. Evaluating the UMLS as a source of lexical knowledge for medical language processing. *Proc AMIA Symp,* pp.189-93.

Gardner, J. and Xiong, L. 2008. HIDE: An Integrated System for Health Information De-identification. *Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems,* pp.254 - 9.

Gardner, J. and Xiong, L. 2009. An integrated framework for de-identifying unstructured medical data. *Data & Knowledge Engineering,* **68**(12), pp.1441-1451.

Girju, R., Badulescu, A. and Moldovan, D. 2006. Automatic Discovery of Part-Whole Relations. *Comput. Linguist.,* **32**(1), pp.83-135.

Guo, Y. 2006. Identifying Personal Health Information Using Support Vector Machines. *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, Washington, DC*.

Gupta, D., Saul, M. and Gilbertson, J. 2004. Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. *Am J Clin Pathol*, pp.176 - 86.

Hahn, U., Romacker, M. and Schulz, S. 2002. Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. *Pac Symp Biocomput*, pp.338-49.

Hanisch, D., Fluck, J., Mevissen, H. T. and Zimmer, R. 2003. Playing biology's name game: identifying protein names in scientific text. *Pac Symp Biocomput*, pp.403-14.

Hara, K. 2006. Applying a SVM Based Chunker and a Text Classifier to the Deid Challenge. *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data, Washington, DC*.

Hina, S., Atwell, E. and Johnson, O. 2010a. Secure Information Extraction from Clinical Documents Using SNOMED CT Gazetteer and Natural Language Processing *In: The 5th International Conference for Internet Technology and Secured Transactions (ICITST-2010)*. IEEE.

Hina, S., Atwell, E. and Johnson, O. 2010b. Semantic Tagging of Medical Narratives with Top Level Concepts from SNOMED CT Healthcare Data Standard. *International Journal of Intelligent Computing Research (IJICR),* **1**(3), pp.118-123.

Hina, S., Atwell, E. and Johnson, O. 2011. Enriching a healthcare corpus with SNOMED CT standard medical semantic tags. *In: Corpus Linguistics 2011, Discourse and Corpus Linguistics, Birmingham*. pp.76-77.

Hina, S., Atwell, E. and Johnson, O. 2012. Automated analysis of domain specific corpus in healthcare domain: For non-domain users. *In: The Sixth Inter-Varietal Applied Corpus Studies (IVACS) International Conference, United Kingdom*.

Hina, S., Atwell, E. and Johnson, O. 2013a. SnoMedTagger: A semantic tagger for medical narratives. *to be published in International Journal of Computational Linguistics and Applications*.

Hina, S., Atwell, E. and Johnson, O. 2013b. SnoMedTagger: A semantic tagger for medical narratives. *In: 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013),* March 24–30, 2013., University of the Aegean, Samos, Greece.

Hina, S., Atwell, E., Johnson, O. and Brierley, C. 2013. Identification, Classification and Anonymisation of 'Protected Health Information' in real-time medical data for research purposes *In: The 23rd Meeting of Computational Linguistics in the Netherlands (CLIN 2013), Netherlands*.

Hina, S., Atwell, E., Johnson, O. and West, R. 2010. Extracting the concepts in Clinical Documents using SNOMED-CT and GATE. *In: Fourth i2b2/VA Shared-Task and Workshop, Challenges in Natural Language Processing for Clinical Data*.

Hirschman, L., Friedman, C., Mcentire, R. and Wu, C. H. 2003. Linking Biomedical Language, Information and Knowledge - Session Introduction. *In: In Proceedings of Pacific Symposium on Biocomputing*, pp.388-390.

Hripcsak, G. and Rothschild, A. 2005. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Informatics Association,* **12**(3), pp.296 - 298.

Huang, Y., Lowe, H., Klein, D. and Cucina, R. 2005. Improved identification of noun phrases in clinical radiology reports using a high-performance statistical natural language parser augmented with the UMLS specialist lexicon. *J Am Med Inform Assoc,* **12**(3), pp.275 - 85.

i2b2: Informatics for Integrating Biology & the Bedside. [online]. [Accessed 12-04-2010]. Available from: http://www.i2b2.org.

International Challenge: Classifying Clinical Free Text Using Natural Language Processing. [online]. [Accessed 30-06-2012]. Available from: http://computationalmedicine.org/challenge/previous

Jonquet, C., Musen, M. A. and Shah, N. H. 2010. Building a biomedical ontology recommender web service. *Journal of Biomedical Semantics,* **1**(1), pS1.

Jurafsky, D. and Martin, J. H. 2009. *Speech and Language Processing* An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recongnition. 2nd ed. New Jersey: Pearson Education.

Kilicoglu, H., Rosemblat, G., Fiszman, M. and Rindflesch, T. 2011. Constructing a semantic predication gold standard from the biomedical literature. *BMC Bioinformatics,* **12**(1), p486.

Kirchner, F. and Sinot, F.-R. 2007. Rule-Based Operational Semantics for an Imperative Language. *Electronic Notes in Theoretical Computer Science (ENTCS),* **174**(1), pp.35-47.

Klebanov, B. B. and Beigman, E. 2009. From annotator agreement to noise models. *Comput. Linguist.*, **35**(4), pp.495-503.

Krauthammer, M., Rzhetsky, A., Morozov, P. and Friedman, C. 2000. Using BLAST for identifying gene and protein names in journal articles. *Gene*, **259**(1-2), pp.245-52.

Leech, G., Garside, R. and Atwell, E. 1983. The Automatic Grammatical Tagging of the LOB Corpus. *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*, **7**, pp.13-33.

Lew, R. and Mitton, R. 2012. Online English Learners' Dictionaries and Misspellings: One Year On. *International Journal of Lexicography*.

Li, Y. and Shawe-Taylor, J. 2003. The SVM with uneven margins and Chinese document categorization. *In: The 17th pacific Asia Conference on Language , Information and Computation (PACLIC17)*, Singapore. pp.216-227.

Lindberg, D. A., Humphreys, B. L. and Mccray, A. T. 1993. The Unified Medical Language System. *Methods of information in medicine*, **32**(4), pp.281-291.

LingPipe 4.1.0. [online]. [Accessed 01-10-2008]. Available from: http://alias-i.com/lingpipe.

Litman, D., Hirschberg, J. and Swerts, M. 2006. Characterizing and Predicting Corrections in Spoken Dialogue Systems. *Comput. Linguist.*, **32**(3), pp.417-438.

Liu, K., Mitchell, K. J., Chapman, W. W. and Crowley, R. S. 2005. Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set. *AMIA Annu Symp Proc*, pp.460-4.

Long, W. 2005. Extracting Diagnosis from Discharge Summaries. *AMIA Annu Symp Proc*, **2005**, pp.470-474.

Marciniak, M., Mykowiecka, A. and Rychlik, P. 2010. Medical text data anonymization. *Journal of medical informatics & technologies*, **16**.

Markert, K. and Nissim, M. 2002. Metonymy resolution as a classification task. *In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*. Association for Computational Linguistics, pp.204-213.

Mccormick, P. J., Elhadad, N. and Stetson, P. D. 2008. Use of Semantic Features to Classify Patient Smoking Status. *Proc AMIA Symp*, pp.450–454.

Mccray, A. T., Aronson, A. R., Browne, A. C., Rindflesch, T. C., Razi, A. and Srinivasan, S. 1993. UMLS knowledge for biomedical language processing. *Bulletin of the Medical Library Association*, **81**(2), pp.184-194.

Meystre, S. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, pp.128 - 44.

Meystre, S., Friedlin, F., South, B., Shen, S. and Samore, M. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, **10**(1), p70.

Meystre, S. and Haug, P. 2005. Automation of a problem list using natural language processing. *BMC Med Inform Decis Mak*, **5**, p30.

Morrison, F. 2009. Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes? *J Am Med Inform Assoc*, **16**(1), pp.37 - 9.

Nadeau, D. and Turney, P. 2005. A Supervised Learning Approach to Acronym Identification. *In: In Proceedings of Canadian Conference on AI'2005*. pp.319-329.

Nadeau, D., Turney, P. and Matwin, S. 2006. Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity. *In:* L. LAMONTAGNE and M. MARCHAND, eds. *Advances in Artificial Intelligence.* Springer Berlin / Heidelberg, pp.266-277.

Neamatullah, I. 2008. Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak*, **8**, p32.

NHS-Connecting for Health. [online]. [Accessed 19-06-2013]. Available from: http://www.connectingforhealth.nhs.uk/systemsandservices/data/uktc/snomed.

Nlm. 2011. *Meaningful Use Quality Performance Measures Benefit from New SNOMED CT "Public Good" Use Policy* [online]. [Accessed 01-03-2011]. Available from: http://www.nlm.nih.gov/news/snomed_perform_measure.html.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G. and Musen, M. A. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research,* **37**(suppl 2), pp.W170-W173.

Ogren, P., Savova, G. and Chute, C. 2008. Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition. *In: LREC*.

Ohta, T., Tateisi, Y. and Kim, J.-D. 2002. The GENIA corpus: an annotated research abstract corpus in molecular biology domain. *In: Proceedings of the second international conference on Human Language Technology Research, San Diego, California*. Morgan Kaufmann Publishers Inc.

Pakhomov, S., Buntrock, J. and Duffy, P. 2005. High throughput modularized NLP system for clinical text. *In: Proceedings of the ACL 2005 on Interactive poster and demonstration sessions, Ann Arbor, Michigan*. Association for Computational Linguistics, pp.25-28.

Patterson, O., Lgo, S. and Hurdle, J. F. 2010. Automatic Acquisition of Sublanguage Semantic Schema: Towards the Word Sense Disambiguation of Clinical Narratives. *In: AMIA Annual Symposium*: AMIA, pp.612-616.

Pestian, J. P., Brew, C., Matykiewicz, P., Hovermale, D. J., Johnson, N., Cohen, K. B. and Duch, W. 2007. A shared task involving multi-label classification of clinical free text. *In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, Prague, Czech Republic*. Association for Computational Linguistics, pp.97-104.

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D. and Goranov, M. 2003. KIM – Semantic Annotation Platform. *In:* D. FENSEL, K. SYCARA and J. MYLOPOULOS, eds. *The Semantic Web - ISWC 2003.* Springer Berlin / Heidelberg, pp.834-849.

Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D. and Kirilov, A. 2004. KIM : a semantic platform for information extraction and retrieval. *Nat. Lang. Eng.*, **10**(3-4), pp.375-392.

Pustejovsky, J., Castaño, J., Cochran, B., Kotecki, M. and Morrell, M. eds. 2001a. *Automatic Extraction of Acronym-meaning Pairs from MEDLINE Databases.* IOS Press.

Pustejovsky, J., Castano, J., Cochran, B., Kotecki, M., Morrell, M. and Rumshisky, A. 2001b. Extraction and disambiguation of acronym-meaning pairs in medline. *Medinfo,* **10**, pp.371-375.

Rau, L. F. 1991. Extracting company names from text. *In: Seventh IEEE Conference on Artificial Intelligence Applications*, 24-28 Feb 1991, pp.29-32.

Roberts A, G. R., Hepple M, Davis N, Demetriou G, Guo Y, Kola J, Roberts I, Setzer a, Trapuria a, Wheeldin B. 2007. The CLEF corpus: semantic annotation of clinical text. *AMIA Annu Symp Proc*, pp.625-629.

Roth, D. and Yih, W.-T. 2002. Probabilistic reasoning for entity & relation recognition. *In: Proceedings of the 19th international conference on Computational linguistics - Volume 1, Taipei, Taiwan*. Association for Computational Linguistics, pp.1-7.

Ruch, P. 2000. Medical document anonymization with a semantic lexicon. *Proc AMIA Symp*, pp.729 - 33.

Sable, J. H., Nash, S. K. and Wang, A. Y. 2001. Culling a clinical terminology: a systematic approach to identifying problematic content. *Proc AMIA Symp*, pp.578-82.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C. and Chute, C. G. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, **17**(5), pp.507-513.

Sawalha, M. and Atwell, E. 2013. A standard tag set expounding traditional morphological features for Arabic language Part-of-Speech tagging. *Word Structure Journal*, **6**, pp.43-99.

Schwartz, A. and Hearst, M. 2003. A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text. *Proceedings of the 8th Pacific Sym-posium on Biocomputing: 03-07 January 2003; Lihue, Hawaii*, pp.451 - 462.

Sevenster, M., Ommering, R. and Qian, Y. 2012. Automatically Correlating Clinical Findings and Body Locations in Radiology Reports Using MedLEE. *Journal of Digital Imaging*, **25**(2), pp.240-249.

Sibanda, T., He, T., Szolovits, P. and Uzuner, O. 2006. Syntactically-informed semantic category recognition in discharge summaries. *AMIA Annu Symp Proc*, pp.714-8.

Skeppstedt, M., Kvist, M. and Dalianis, H. 2012. Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text. *In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 May, Istanbul, Turkey. European Language Resources Association (ELRA).

Sleator, D. and Tamperley, D. C. SCIENCE. 1991. *Parsing English with a Link Grammar*. (CMU-CS-91-196). Pittsburgh: Carnegie Mellon University

Smith, R., Xu, J., Hina, S. and Johnson, O. 2013. GATEway to the Cloud Case Study: A privacy-aware environment for Electronic Health Records research. *In: IEEE MobileCloud2013 Industry Track in conjunction with SOSE 2013*, 25th March - 28th March 2013, San Francisco Bay, USA.

SNOMED CT User Guide, January 2011 International Release. International Health Terminology Standards Development Organisation.

Snyder, B. and Palmer, M. 2004. The English all-words task. *In: Senseval-3: 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, July, Barcelona, Spain. Association for Computational Linguistics, pp.41-43.

Sokolova, M. and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, **45**(4), pp.427-437.

Song, D., Chute, C. and Tao, C. 2011. Semantator: a semi-automatic semantic annotation tool for clinical narratives. *In: 10th International Semantic Web Conference (ISWC2011), Bonn, Germany*.

Song, D., Chute, C. G. and Tao, C. 2012. Semantator: annotating clinical narratives with semantic web ontologies. *AMIA Summits Transl Sci Proc*, **2012**, pp.20-9.

Stearns, M. Q., Price, C., Spackman, K. A. and Wang, A. Y. 2001. SNOMED clinical terms: overview of the development process and project status. *Proc AMIA Symp*, pp.662–666.

Sweeney, L. 1996. Replacing personally-identifying information in medical records, the Scrub system. *Proc AMIA Annu Fall Symp*, pp.333-7.

Szarvas, G., Farkas, R. and Busa-Fekete, R. 2007. State-of-the-art anonymization of medical records using an iterative machine learning framework. *J Am Med Inform Assoc*, pp.574 - 80.

Szarvas, G., Farkas, R. and Kocsor, A. 2006. A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms. *9th Int Conf Disc Sci (DS2006), LNAI*, pp.267 - 278.

Taira, R., Bui, A. and Kangarloo, H. 2002. Identification of patient name references within medical documents using semantic selectional restrictions. *Proc AMIA Symp*, pp.757 - 61.

Taira, R. K. and Soderland, S. G. 1999. A statistical natural language processor for medical reports. *Proc AMIA Symp*, pp.970-4.

Tang, B., Cao, H., Wu, Y., Jiang, M. and Xu, H. 2013. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. *BMC Medical Informatics and Decision Making*, **13**(Suppl 1), pS1.

Thomas, S. M., Mamlin, B., Schadow, G. and Mcdonald, C. 2002. A successful technique for removing names in pathology reports using an augmented search and replace method. *Proc AMIA Symp*, pp.777-81.

Thompson, P., Iqbal, S., Mcnaught, J. and Ananiadou, S. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, **10**(1), p349.

Turchin, A., Kolatkar, N. S., Grant, R. W., Makhni, E. C., Pendergrass, M. L. and Einbinder, J. S. 2006. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *Journal of the American Medical Informatics Association : JAMIA*, **13**(6), pp.691-695.

Tveit, A., Edsberg, O., Røst, T. B., Faxvaag, A., Nytrø, Ø., Nordgård, T., Ranang, M. T. and Grimsmo, A. 2004. Anonymization of General Practitioner Medical Records. *In: second HelsIT Conference at the Healthcare Informatics week* Trondheim.

Unified Medical Language System® (UMLS®). [online]. [Accessed 04-04-2010]. Available from: http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html.

Uzuner, Ö., Goldstein, I., Luo, Y. and Kohane, I. 2008a. Identifying Patient Smoking Status from Medical Discharge Records. *Journal of the American Medical Informatics Association,* **15**(1), pp.14-24.

Uzuner, O., Luo, Y. and Szolovits, P. 2007. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc,* **14**(5), pp.550-63.

Uzuner, Ö., Sibanda, T. C., Luo, Y. and Szolovits, P. 2008b. A de-identifier for medical discharge summaries. *Artificial Intelligence in Medicine,* **42**(1), pp.13-35.

Uzuner, Ö., South, B. R., Shen, S. and Duvall, S. L. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*.

Vieira, R. and Poesio, M. 2000. An empirically based system for processing definite descriptions. *Comput. Linguist.,* **26**(4), pp.539-593.

Wain, H. M., Lush, M., Ducluzeau, F. and Povey, S. 2002. Genew: the Human Gene Nomenclature Database. *Nucleic Acids Research,* **30**(1), pp.169-171.

Wang, X. 2007. Rule-Based Protein Term Identification with Help from Automatic Species Tagging. *In: Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing, Mexico City, Mexico*. Springer-Verlag, pp.288-298.

Wellner, B. 2007. Rapidly retargetable approaches to de-identification in medical records. *J Am Med Inform Assoc,* pp.564 - 73.

Whetzel, P. and Team, N. 2013. NCBO Technology: Powering semantically aware applications. *Journal of Biomedical Semantics,* **4**(Suppl 1), pS8.

Yu-Chieh Wu, Teng-Kai Fan, Yue-Shi Lee and Yen, S.-J. 2006. *Extracting Named Entities Using Support Vector Machines. KDLL'06.*

Zeng, Q., Goryachev, S., Weiss, S., Sordo, M., Murphy, S. and Lazarus, R. 2006. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making,* **6**(1), p30.

Zhou, M. and Huang, C. 1994. An efficient syntactic tagging tool for corpora. *In: Proceedings of the 15th conference on Computational linguistics - Volume 2, Kyoto, Japan*. Association for Computational Linguistics, pp.949-955.