EXTENDING THE HIRST AND ST-ONGE MEASURE OF SEMANTIC RELATEDNESS FOR THE UNIFIED MEDICAL LANGUAGE SYSTEM

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Mugdha Choudhari

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Dr. Ted Pedersen

August, 2012

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Prof. Ted Pedersen for the continuous support and understanding during my Masters study and research. His guidance and reachability through emails, helped me in all times of research and writing of this thesis.

I would like to express my appreciation to my advisory committee: Dr. Hudson Turner and Dr. Marshall Hampton for their insightful comments and feedback. Special thanks to Dr. Maclin for his guidance through stimulating discussions, help and understanding. Also, thanks to the Department of Computer Science for making this study possible. My gratitude also goes to all my professors along with Dr. Crouch, for their support and guidance.

I thank my fellow classmates and friends for making my Masters the most memorable time of life and for patiently listening to me in my good and bad times. Furthermore, I would also like to thank the Computer Science Office staff, for their help.

The most special thanks goes to my best partner and friend, my husband. Ajay, you gave me your unconditional support, motivation and love through all this process.

My deepest gratitude goes to my family, my parents and parents-in-law for always standing behind me and for giving their valuable support and love.

Last but not the least, thank you God, for the strength you have always given me. I am grateful to the beautiful city of Duluth and the huge calm Lake Superior for being my source of peace.

Abstract

One of the organizational structures of human memory is clustering, i.e., related concepts are usually stored together. This is one of the reasons why humans are good at determining how closely related two concepts are to each other. As a large number of natural language processing applications use semantic relatedness, we need to give machines the ability to calculate how closely related two concepts are. In past decades, different researchers have taken approaches to solve this problem by either using the context of the concept or using the concept network structures where concept are denoted as nodes and relationships between them as edges. Hirst and St-Onge (HSO) presented a measure of semantic relatedness using WordNet, by introducing the concept of an allowable path. In this thesis we address the problem of finding semantic relatedness between two biomedical concepts by applying HSO measure on Unified Medical Language System (UMLS), a graph of biomedical concepts connected with different medical relations.

The HSO measure is applied to UMLS and evaluated using Spearman's Correlation Coefficient against gold standards based on human judgments for different experimental data sets. We observe that the structure of UMLS is very different than WordNet, as UMLS is larger in size and has denser connections. It is also proven that the cost of each horizontal link should be greater than each up or down link for achieving better correlation, as suggested by HSO algorithm. The experimental evidence shows that the correlation values of HSO measure with upward and downward links are comparable to path measure with UMLS::Similarity. Addition of experimentally chosen horizontal relations and attribute leads to an improvement in the correlation values. To perform the experiments with large set of chosen horizontal relations and attributes, we present big subsets of the experimental data sets. Furthermore, restricting the path length in each direction provides further improvement to the correlation with gold standards. We also find that allowing two direction changes instead of one (as in original HSO) leads to better correlation in most of the cases. We conclude that, the HSO measure can be extended to accommodate the UMLS structure and to find semantic relatedness between concepts of biomedical domain.

Contents

Acknowledgements **i**

Abstract **ii**

List of Tables **vi**

List of Figures **ix**

1 Introduction **1**

1.1 Motivation 1

1.2 Thesis' Methodology 3

1.3 Contributions of the thesis 5

2 Background **8**

2.1 Hirst and St-Onge (HSO) Measure 8

2.1.1 Cohesive Relations 9

2.1.2 Allowable Paths 11

2.1.3 Formulation of HSO measure 13

2.2 WordNet 14

2.3 Unified Medical Language System 15

2.3.1 UMLS Metathesaurus 15

2.3.2 SNOMEDCT vocabulary 17

2.3.3 MSH vocabulary 25

2.3.4 Accessing UMLS Data 27

3 Algorithm **29**
3.1 Accessing UMLSKS using Web Services 30
3.1.1 Authenticating User 31
3.1.2 Verification of Input 31
3.1.3 Accessing data 31
3.2 Determining value of constants C and k in HSO formulation 34
3.3 Calculating Semantic Relatedness 36
3.3.1 Graph Formation 37
3.3.2 Finding Shortest Allowable path 41
3.4 Implementation and Configuration Details 45

4 Experimental Data **47**
4.1 Rubenstein and Goodenough's Data 48
4.2 MayoSRS and MiniMayoSRS data sets 48
4.3 UMNSRS_reduced_rel and UMNSRS_reduced_sim test sets 51
4.4 MiniMayoSRS set for MSH vocabulary 52
4.5 Big subsets for SNOMECT 53
4.6 Spearman's Rank Correlation Coefficient Evaluation 53

5 Experimental Results **58**
5.1 Hypothesis 1: The HSO measure when applied with only up (PAR relation) and down (CHD relation) vectors is equivalent to the shortest path measure implemented by the UMLS::Similarity package. 61
5.2 Hypothesis 2: All relations attributes from SNOMEDCT vocabulary from relation RO (other relations) can be used to represent horizontal links. . 64
5.3 Hypothesis 3: The addition of horizontal relations and attributes selected by hypothesis 2 improves the correlation to the gold standards. 69
5.4 Hypothesis 4: When the cost of traveling one Horizontal link is greater than the cost of one vertical link, the correlation to the gold standards is improved. 76
5.5 Hypothesis 5: All possible allowable path patterns described by the HSO measure can be observed in the SNOMEDCT vocabulary, as it is a sufficiently large vocabulary. 82

5.6 Hypothesis 6: If the path vectors in an allowable path are restricted in length, it correlates more with gold standard values, as it reduces the number of false positives. 90
5.7 Hypothesis 7: Allowing two direction changes in an allowable path between medical concepts, aids in improving the correlation with gold standards. 94

6 Related Work **98**
6.1 Application of HSO for Malapropism Detection 98
6.2 Development of a conceptual distance metric for the UMLS 99
6.3 Measures of semantic similarity and relatedness in the biomedical domain 100
6.4 Comparison of Ontology-based Semantic- Similarity Measures 101
6.5 UMLS::Similarity 102

7 Conclusion **103**

8 Future Work **105**

References **107**

Appendix A. Additional Results **110**

List of Tables

2.1 Major sources in UMLS Metathesaurus 17
2.2 Most Frequent Relation Attributes in SNOMEDCT 19
3.1 MiniMayoSRS Correlation Values for different values of constant k . . . 36
4.1 MiniMayoSRS test set 50
4.2 MiniMayoSRS.msh test set 52
4.3 MiniMayoSRS Big subset test set with key 54
4.4 Spearman's Rank Assignment 55
4.5 Spearman's Correlation Coefficient Example 56
4.6 Spearman's Correlation Coefficient Calculation 56
5.1 Baseline Spearman's Correlation for SNOMEDCT and MSH (REL:PAR, DIR:U) 62
5.2 Comparison between correlation values of HSO default configuration (SAB: SNOMEDCT) and UMLS-Similarity path measure 62
5.3 Comparison between correlation values of HSO default configuration (SAB:MSH) and UMLS-Similarity path measure 63
5.4 Set of CUIs from MiniMayoSRS (SAB : SNOMEDCT, REL : PAR,RO, DIR : U,H) 65
5.5 Top Attributes' from RO relation 67
5.6 Correlation values using default configuration (SAB : SNOMEDCT) . . 71
5.7 Number of CUI pairs for which SR value increased after adding H links using cost of H - 2 (SAB : SNOMEDCT) 71
5.8 Comparison between correlation values of default configuration and configuration with H relations, cost of H - 2 (SAB : SNOMEDCT) 72
5.9 Spearman's Correlation Values (SAB:MSH, REL:PAR,SIB DIR:U,H) . . 74

5.10 Comparison between correlation values of default configuration and Configuration with H relations, cost of H - 1 (SAB : SNOMEDCT) 77
5.11 Number of CUI pairs for which path with H link was replaced by shorter path with U/D links, cost of H - 3 (SAB : SNOMEDCT) 79
5.12 Comparison between correlation values of default configuration and Configuration with H relations, cost of H - 3 (SAB : SNOMEDCT) 79
5.13 Comparison between number of CUI pairs for which SR value increased after adding H links using cost of H link is 3 and 2 (SAB : SNOMEDCT) 80
5.14 Comparison between semantic relatedness values after adding H links using cost of H link is 3, 2 and 1 (SAB : SNOMEDCT) 80
5.15 Comparison between number of CUI pairs for which SR value increased after adding H links using cost of H link is 3, 2 and 1 (SAB : SNOMEDCT) 80
5.16 Spearman's Correlations after restricting vector length (l) to 4, 5, 6 and 8 using default configuration(Figure 5.1)(SAB : SNOMEDCT) 92
5.17 Spearman's Correlations after restricting vector length to 4, 5, 6 and 8 using default configuration(Figure 5.2)(SAB : MSH) 92
5.18 Comparison between semantic relatedness with and without path restriction after adding H links (SAB : SNOMEDCT) 93
5.19 Comparison between correlation values with one direction (1D) and two direction (2D) changes allowed using H relations (SAB : SNOMEDCT and MSH) 96
A.1 MayoSRS test set (part 1) 111
A.2 MayoSRS test set (part 2) 112
A.3 MayoSRS test set (part 3) 113
A.4 MiniMayoSRS set (SAB:SNOMEDCT, REL:PAR, DIR:U) 114
A.5 MayoSRS set 1 (SAB : SNOMEDCT, REL : PAR, DIR : U) 115
A.6 MayoSRS set 2 (SAB : SNOMEDCT, REL : PAR, DIR : U) 116
A.7 MayoSRS set 3 (SAB : SNOMEDCT, REL : PAR, DIR : U) 117
A.8 MayoSRS Big subset 1 test set with key 118
A.9 MayoSRS test set Big subset 2 with key 119
A.10 MiniMayoSRS big subset (SAB:SNOMEDCT, REL:PAR, DIR:U) 120
A.11 MayoSRS Big subset 1 (SAB : SNOMEDCT, REL : PAR, DIR : U) . . 121

A.12 MayoSRS Big subset 2 (SAB : SNOMEDCT, REL : PAR, DIR : U) . . 122
A.13 MiniMayoSRS big subset (SAB:SNOMEDCT, REL : PAR,RB,RN,RO, DIR:U,H,H,H) Cost of H link is 3 123
A.14 MayoSRS Big subset 1 (SAB : SNOMEDCT, REL : PAR,RB,RN,RO, DIR : U,H,H,H), Cost of H link is 3 124
A.15 MayoSRS Big subset 2 (SAB : SNOMEDCT, REL : PAR,RB,RN,RO, DIR : U,H,H,H), Cost of H link is 3 125
A.16 MiniMayoSRS big subset (SAB:SNOMEDCT, REL : PAR,RB,RN,RO, DIR:U,H,H,H) Cost of H link is 2 126
A.17 MayoSRS Big subset 1 (SAB : SNOMEDCT, REL : PAR,RB,RN,RO, DIR : U,H,H,H), Cost of H link is 2 127
A.18 MayoSRS Big subset 2 (SAB : SNOMEDCT, REL : PAR,RB,RN,RO, DIR : U,H,H,H), Cost of H link is 2 128
A.19 MiniMayoSRS big subset (SAB:SNOMEDCT, REL : PAR,RB,RN,RO, DIR:U,H,H,H) Cost of H link is 1 129
A.20 MayoSRS Big subset 1 (SAB : SNOMEDCT, REL : PAR,RB,RN,RO, DIR : U,H,H,H), Cost of H link is 1 130
A.21 MayoSRS Big subset 2 (SAB : SNOMEDCT, REL : PAR,RB,RN,RO, DIR : U,H,H,H), Cost of H link is 1 131

List of Figures

2.1 Allowable and Non-allowable paths example 11
2.2 Allowable Patterns 12
2.3 Relations in UMLS Metathesaurus 18
2.4 Defining attributes used for defining Bacterial Pneumonia 20
3.1 Sample query of findCUIByExact in Perl 32
3.2 Sample output of findCUIByExact 33
3.3 Sample query of getConceptProperties in Perl 34
3.4 Sample configuration file 46
3.5 Default HSO allowable patterns regular expression 46
5.1 Default SNOMEDCT configuration file 61
5.2 Default MSH configuration file 62
5.3 Selected H relations and attributes 68
5.4 Configuration file with selected H relations and attributes 69
5.5 Examples with updated path after adding H relations 73
5.6 MSH configuration file with SIB relation as H link 74
5.7 Path between C0002962(Angina Pectoris) and C0070166(clopidogrel) 78
5.8 Allowable path pattern 1 and 2 84
5.9 Allowable path pattern 3 and 4 85
5.10 Allowable path pattern 5 and 6 86
5.11 Allowable path pattern 7 87
5.12 Allowable path pattern 8 88
5.13 Path between Cholangiocarcinoma (C0206698) and Colonoscopy (C0009378) 90
5.14 HSO allowable patterns' regular expression after restricting vector length to 4 91

5.15 HSO allowable patterns' regular expression after restricting vector length to 5 91
5.16 HSO allowable patterns' regular expression after restricting vector length to 6 91
5.17 HSO allowable patterns' regular expression after restricting vector length to 8 92
5.18 Patterns with 2 direction changes allowed 94
5.19 Path between C0333997(Lymphoid hyperplasia) and C0007107(Malignant neoplasm of larynx) 95

Chapter 1

Introduction

In this chapter, we introduce the problem of finding semantic relatedness between medical concepts. We describe motivation behind solving this problem and the thesis's methodology of extedning the Hirst and St-Onge measure of semantic relatedness to the Unified Medical Language System.

1.1 Motivation

How much does the term **clock** have to do with **time**? The answer would be that they have a strong relation. Semantic relatedness between two concepts tells us how strongly they are connected with each other. In other words, what comes to mind when we think of *eyes* may be *vision*, *cornea*, *organ*, *pupil* etc. All these words are semantically related to *eyes* by different relations. The nature of relation between the terms can be synonymy (have similar meaning), antonymy (have opposite meaning), meronymy (part-name), hyponymy (sub-name), functional, associative and others. For example, *inhale* and *exhale* are closely related as they are exactly opposite actions.

Humans are naturally very good at relating concepts to one another. For example, it is easier for a person to tell that *Schizophrenia* is more closely related to *mental disability* than it is related to *muscle sprain*. On the other hand it is not so easy for machines to understand or tell how strongly/weakly two concepts are related. Efforts have been made for several years, to make machines capable of identifying how similar or different given two concepts are. Solving this problem is important as it is useful in areas

of information retrieval, automatic text correction, speech recognition, summarization, search engines and many more. For example, while retrieving the document based on occurrence of search query term *aorta*, we can also retrieve documents that contain words from the set of semantically related words to *aorta* such as *heart*, *blood*, *artery*, *vein*, *blood vessel*, *circulation*, etc. But all the words in related words set do not share equal relatedness with *aorta*. Some of them are more related than others. The measure of semantic relatedness can be used to quantify how much each of the terms in related set is semantically related to *aorta*. The measure assigns a numerical score that would define the degree to which each term from the set is related to query term (in this example *aorta*). Thus a pairwise calculation of semantic relatedness can be used to prioritize the search results.

Semantic similarity and semantic distance are interchangeably used in literature for semantic relatedness. Semantic similarity is a specific case of semantic relatedness, that justifies how similar two concepts are. Two concepts are semantically similar if one concept shares 'isa' relation with another. For example, *heartburn* is semantically 'similar' to *burning reflux*, as they share common properties and *heartburn* 'isa' *burning reflux*, whereas *heartburn* is semantically 'related' to *esophagus*. Semantic distance, on the other hand, can be thought inverse of semantic similarity. The more the semantic distance between two concepts, the less they are semantically similar. For example, as *fat(substance)* and *nail plate*, have a high semantic distance between them, thus semantic similarity between them is very low.

Measuring the semantic relatedness between two words from biomedical domain falls under specific area of this problem of finding semantic relatedness and is useful in the field of Bioinformatics. Measuring the relatedness of words requires real-world knowledge about entities and concepts which is difficult to obtain from meaning of just the given pair of words. Fortunately there are some constructed knowledge sources such as UMLS and WordNet that can be used in the task of measuring semantic relatedness. These knowledge sources provide semantic networks in which nodes represent the concepts and the arcs joining them represent the relation by which the concepts are related.

The Unified Medical Language System (UMLS) is an on-line database of biomedical and health data from different available sources. It provides knowledge sources and

tools which can be used by developers to build systems for different purposes related to public health or applications in the fields of Bio-informatics and Natural Language Processing [1]. Similar to UMLS, WordNet is a large lexical database of English words and is widely used by researchers to solve different language processing problems.

The problem of quantifying semantic relatedness between two concepts has a long history in artificial intelligence, philosophy and psychology, going back to Aristotle(384-322 B.C.E.) [2]. Different researchers such as Osgood [3], Quillian [4], Collins and Loftus(1975) [5] applied different approaches to solve the problem of measuring semantic relatedness. Osgood tried to represent words as entities in n-dimensional space and measured the distance between them using the rules of Euclidean geometry, the technique known as 'semantic differential'. Quillian and Collins and Loftus, worked on a procedural approach known as 'spreading activation' or 'marker parsing', which was based on the idea that context of the word indicates the semantic distance. The work also suggested that measures of semantic distance are inherent in network structure, assuming concepts form the nodes of the network. The work was further continued by Hirst(1987) [6].

This thesis mainly focuses on finding the semantic relatedness, sometimes known as semantic distance, between concepts in the biomedical domain. It is important to find semantic relatedness between biomedical concepts as it helps in finding medical articles with similar content, in searching the patients with similar diseases from their medical reports and then this information can be used for medical surveillance. It will also help in solving the problems faced in the organization of biomedical terminologies and ontologies.

1.2 Thesis' Methodology

This thesis presents an approach of measuring semantic relatedness between two medical concepts by extending the measure proposed by Hirst and St-Onge (HSO) in 1995 [6]. Hirst and St-Onge's measure was originally applied to WordNet which is a smaller graph as compared to UMLS and has limitations in the set of relations along with inconsistency in links. As UMLS has large number of medical concepts from different sources integrated consistently, application of HSO on UMLS data leads to interesting

results.

Previously developed measures of semantic relatedness use the knowledge of path distance between two medical concepts but give less attention to changes in direction of the path. Whereas Hirst and St-Onge's measure takes into account the direction of the traversal path in network by introducing the concept of allowable path. A path is allowed if it will lead to a destination concept semantically related to the source concept. To decide if the path is allowable or not allowable, a set of allowable path patterns (with another set of non-allowable patterns) consisting of all the path patterns which do not digress from the context of original concept is provided by HSO measure. The path patterns are composed of connected vectors in up, down and horizontal directions. If the two concepts under consideration are connected to each other by one of the allowable paths from this set, then this path is considered for the calculation of semantic relatedness. Hirst and St-Onge's original work is described in detail as it provides a necessary background of this thesis work. UMLS's structure and source vocabularies such as SNOMEDCT and MSH are also discussed in depth as it provides us with the medical thesaurus required for measuring the semantic relatedness.

A Breadth First Search (BFS) and Dijkstra based algorithm customized to accommodate HSO's approach is adapted to find the semantic distance between two concepts using UMLS meta-thesaurus knowledge source. The algorithm uses the concept of allowable path to effectively find an allowable shortest path between two medical concepts and has been implemented in both PERL and JAVA to perform various kinds of experiments.

The work is extensively evaluated by using the Spearman's Correlation Coefficient and the results are compared with gold standards (formed by human judgments) on various experimental data sets. UMLS is accessed using Web services which avoids the drawbacks of having complete UMLS database on local machine and reduces inconsistency in data.

To study the effect of extending HSO measure to UMLS data, we present a list of hypotheses which verify HSO's original suggestions and findings against medical concepts. We then perform a series of experiments in an attempt to prove these hypothesis and then analyze the experimental results.

1.3 Contributions of the thesis

Here are some major findings of this thesis work, during the process for extension of HSO measure to UMLS and accommodating it for medical concepts:

- It was observed that the structure of UMLS is very different than that of WordNet, as UMLS is larger in size and has denser connections/relationships between the concepts. This difference affects the performance of HSO measure when applied to UMLS.

- According to HSO measure, the paths which have higher number of changes in direction should be penalized for each direction change they make. We penalize an allowable path by decaying the value of semantic relatedness for each direction change made by the path. Out experimental results show that semantic relatedness values correlate with human judgments by following this suggestion.

- We confirm that using the concept of allowable paths with up and down links, results in semantic relatedness values that have satisfying agreement with human judgments. The correlation values are functionally equivalent to those using path measure with UMLS::Similarity. Thus, using HSO with up and down relations, we are able to find results similar the shortest path measure.

- We observe that many concepts in SNOMEDCT (one of the sources in UMLS), have very large number of connected concepts related by child (CHD) relation or by other relations (RO). This makes it difficult to access such concepts over the network and to operate on them efficiently. We handle this issue by forming the big subsets of experimental data sets with all the concepts that have manageable number of connected concepts.

- We find that not all the horizontal relations and attributes in UMLS yield meaningful paths between medical concepts and thus we filter the top relations and attributes, thus presenting a set of attributes that help in improving the correlation with human judgments.

- We show that using horizontal links to find a path between two concepts results in interesting path patterns (both allowed and not allowed) and also leads to an

improvement over correlation values obtained by just using up and down links.

- Allowing unrestricted extension of allowable path vector in each direction sometimes relates the concepts which are at high semantic distance, thus lowering the agreement with human judgments. We attempted to restrict the length of allowable pattern vector in a particular direction and found a considerable improvement over the correlation values obtained by following original HSO allowable patterns set.

- We show that the cost of each horizontal link should be greater than each up or down link while calculating the path distance between concepts, as suggested by HSO algorithm.

- We explore the path patterns found in SNOMEDCT vocabulary and find that all allowable path patterns from HSO's allowable patterns set can be observed in UMLS's SNOMEDCT vocabulary graph. It was also observed that UMLS structure has new path patterns (except from those already present in HSO's set) that can be good candidates for an allowable path patterns set.

- We also study the effect of applying HSO formulation by allowing two direction changes in an allowable path and find that it does lead to improvement in Spearman's correlation values.

Other contributions made by this thesis work are:

- UMLS is accessed through the Web services using SOAP-Lite package in PERL implementation and through UMLSKS API in JAVA implementation. This improves the usability of the measure and reduces the overhead of storing huge UMLS database on local machine.

- We developed a tool (part of PERL package) that can be used to retrieve the definitions and other information about a particular medical concept in UMLS meta-thesaurus.

- We have released (via the CPAN archive) a freely available software package Webservice::UMLSKS::Similarity, which is a PERL implementation of HSO measure

of semantic relatedness for medical concepts.[1] .

- We have also released (via a Sourceforge project) an open source software package Webservice::UMLSKS::Similarity::Java, which is a Java implementation of HSO measure of semantic relatedness for medical concepts.[2]

In brief, this thesis puts forth variety of interesting results, observations along with challenges handled in extending measure of semantic relatedness for medical concepts developed by HSO.

[1] http://search.cpan.org/dist/WebService-UMLSKS-Similarity/
[2] https://sourceforge.net/p/umlsks-sr-java

Chapter 2

Background

To understand application of HSO measure on the Unified Medical Language System concepts' graph, it is necessary to have a relevant background about HSO measure and UMLS's structure. The initial part of this chapter describes original HSO measure along with basic concepts such as *Cohesive relations* and *Allowable paths*. In the later part, we explain UMLS's structure in detail along with its major source vocabularies, relationships and knowledge sources.

2.1 Hirst and St-Onge (HSO) Measure

If there is a close relation between meanings of two concepts or words, then the concepts are said to be semantically related to each other. Semantic relatedness can be quantified by using a measure that determines how closely related two concepts are. Thus, a correct measure of semantic relatedness would assign a higher relatedness value to concepts which are semantically near from each other. Hirst and St-Onge developed one such measure of semantic relatedness and applied the measure to English vocabulary concepts, using WordNet, a lexical database [6]. The measure can be applied to any vocabulary graph which consists of nodes that represent concepts and connections betweens nodes represent different relationships. HSO measure calculates relatedness between concepts using the path distance between the concept nodes, number of changes in direction of the path connecting two concepts and the allowableness of the path. An Allowable Path is a path that does not digress away from the meaning of the source

concept and thus should be considered for the calculation of relatedness. Allowable path patterns were developed by HSO such that different semantic relations such as synonymy, hyponymy, meronymy, association, etc are explored and are accurately used to define the semantic relatedness between the concepts. As we have now outlined the HSO measure in brief, let use now understand each of its key characteristics in detail.

2.1.1 Cohesive Relations

Cohesive relations denote the semantic relationships between concepts, independent of the structure of the cohesive text from which the concepts are obtained. In text with good flow (cohesive text), it is observed that concepts in the current sentence tend to refer to concepts that had already appeared in previous sentences. Such concepts share a cohesive relation of 'identity of reference'[6]. Two concepts can also be related by other cohesive relations such as hyponymy or meronymy or a general association. Such relations between concepts, are independent of the structure of complete text. For example, in (1), the concepts that are bold faced are related to each other by identity of reference, as the word **It** refers to the **weather**. Whereas, in example (2), the bold faced concepts share a hyponymy as **Garlic** is a kind of **Spice** and in example (3), highlighted concepts share a general association of ideas, as **plates**, **spoons**, **bowls** and **table** are associated to each other.

- (1) The **weather** is very unpredictable in Duluth. **It** sometimes gets colder even in Summer. **It** surprises people by contradicting the weather forecasts.
- (2) The most important **spice** in Chicken Curry is **Garlic**.
- (3) Before the dinner, wash the **plates** and **spoons**. Clean the **bowls** and arrange the dinner **table**.

Thus, semantic relations which are independent of the form of the text are cohesive relations. As the basic idea of HSO measure is to determine the semantic relatedness between two concepts using context of the text without complete knowledge of the anatomy of text, it uses different cohesive relations for calculating an allowable path between concepts.

Types of Cohesive Relations Words/concepts in the corpus can be linked to each other by various relations. Hirst and St-Onge introduced three types of cohesive relations which directly correlate to the semantic relatedness between the words [6]. The relations are :

- Extra Strong Relation
- Strong Relation
- Medium Strong Relation

An extra strong relation exists between a word and it's literal repetition. For example, two separate occurrences of *organism* share an extra strong relation between them. Such relations have the highest weight of all relations and result in a high relatedness value.
A strong relation exists between two words which have common parent word or which derive from common parent. For example, *oak* and *pine* are strongly related to each other as they have common parent *tree*. The parent concept for a given word is a concept that is related by *is-a* relation. Two words are said to be strongly related in following cases:

- When the two words share a common parent concept.
- When there exists an association relation like an antonymy or a horizontal link between the parents of the words. For example, *ice* and *steam* are strongly related as they their parent concepts *cold substance* and *hot substance* are antonyms of each other.
- When there is any kind of link at all between a parent of each word if one word is a compound word or a phrase that includes the other. For example, words *color* and *water-color* are strongly related.

An allowable path is a path joining a source word to another word which does not digress away from the meaning of the source word. For example, the path joining *oak* and *leaf* should be allowed as *paper* has close relation to *oak* (paper is made from oak's

bark), whereas a path joining *oak* and *ink* should not be allowed as they are not closely related to each other. Two words are said to be related by a medium strong relation if they are connected by one of the allowable paths from the set of allowable paths. For example, the relation between *oak* and *leaf* is a medium strong relation. It is defined using the allowable path distance between the words along with the direction of the path connecting the two words in the tree of corpus.

The concept of Allowable Path is used by HSO mainly to find semantic relatedness between concepts that share a medium strong relation.

2.1.2 Allowable Paths

The HSO measure states that a concept is semantically far away from another concept if they have a large path distance between them along with large number of changes in direction of the path. A source concept in *WordNet* is connected to other concepts with different paths. Some of these paths digress from the context of the source concept and therefore will lead to a destination concept that is semantically distant from the source concept. HSO avoids such paths so that only the paths which will lead to a destination concept semantically related to the source concept are allowed.

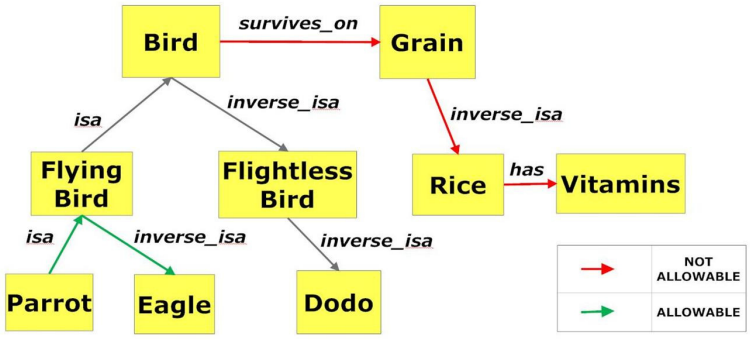

Figure 2.1: Allowable and Non-allowable paths example

In example shown by Figure 2.1.2, *Parrot* and *Eagle* are strongly related to each other and thus the path connecting these two concepts is allowed by HSO measure. But, *Bird* and *Vitamins* are semantically far away from each other and hence are connected

by a path which is not allowed by HSO measure. It makes sense to discard the path connecting *Bird* and *Vitamins* as it is obvious that these two concepts are a lot different from each other. To decide which paths may lead to a semantically related destination concept, HSO defines a set of allowable paths. This is a set of vectors which represent the paths between two concepts. Thus set of allowable paths consists of all the paths which do not digress from the context of original concept. If the two concepts under consideration are connected to each other by one of the allowable paths from this set, then this path is considered for the calculation of weight assignment. Once an allowable path is found, HSO states that the semantic relatedness also depends on the number of changes in the direction of the path. Thus the set of allowable paths gives the patterns that are allowed taking into account the directions and the path distances. HSO also gives the set of path patterns which should not be allowed. Following are the allowable paths for medium strong relation between concepts. Each vector in the Figure 2.1.2 denotes one or more links in the respective direction [6].

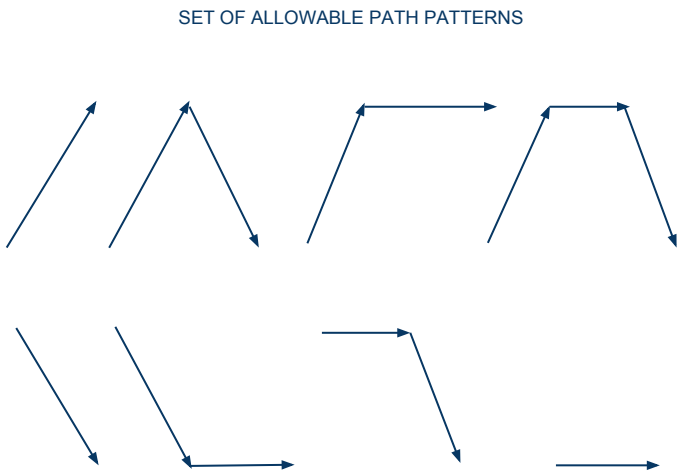

Figure 2.2: Allowable Patterns

HSO measure gives an explanation for why only patterns shown in Figure 2.1.2 are allowed. As shown in example from Figure 2.1.2, an upward link from the concept

symbolizes a generalization. *Flying Bird* is a generalized concept for *Parrot* and *Bird* is a generalized concept for *Flying Bird*. So when consecutive upward links are traversed, the context of the original concept gets generalized. Similarly, as consecutive downward links are followed, the context of the original concept is narrowed down to the specific concept. In both the upward and downward links, the concepts do not digress a lot from the meaning of the original concept. For example, *Parrot*, *Eagle*, *Flying Bird*, *Flightless Bird* and *Bird* are all related to the concept *Bird*. But, as horizontal links which correspond to relations like aggregation, associations, etc., are followed, the concepts tend to digress from the meaning of original concept. From Figure 2.1.2, *Bird* is related to *Grain* by the horizontal relation *survives-on*, *Rice* is a specific *Grain* and *Rice* is related to *Vitamins* by a horizontal *has* relation. This path from *Bird* to *Vitamins* has two horizontal paths. Thus it is observed that the horizontal paths digress a lot from meaning of original concept of *Bird*. HSO define two rules to ensure that a reasonable path exists between source and destination concept.

- Rule 1: No other link precedes an upward link.
  Once the context has been narrowed down by a downward or horizontal link, it is not allowed to generalize the context again by following an upward link.

- Rule 2: At most one change of direction is allowed.
  As we change the direction, it causes a large semantic step, so we should limit the number of changes in direction. However, there is following exception to this second rule:
  It is permitted to use a horizontal link after you follow an upward or downward link.

2.1.3 Formulation of HSO measure

HSO measure formalized the relation between semantic relatedness, the allowable path distance and number of changes in direction of the path between the words, by introducing a scoring mechanism for relatedness. Using the following formula, we can assign a numerical score or weight to the semantic relatedness between two concepts using the details of the path joining them.

$$weight = C - path\ length - k * number\ of\ changes\ of\ direction$$

Here, the path stands for an allowable path between concepts, C and k are constants whose values are derived through experiments.

To conclude, HSO measure gives a good formalization that relates the semantic relatedness between two concepts to the path distance between them and number of direction changes in path. As HSO measure uses a vocabulary graph to find the path between concepts, it can be easily applied to domains other than English vocabulary, given that we have a vocabulary graph available for that domain. Unified Medical Language System (UMLS) is a huge graph of medical concepts and is formed by combining different medical source vocabularies. Even though UMLS's structure is bigger and different than WordNet, by applying HSO on UMLS we can observe and compare the performance of measure on two different vocabulary graphs. We attempt to find semantic relatedness between medical concepts using HSO measure and find its correlation against gold standards formed by human judgments.

2.2 WordNet

HSO originally applied the measure of semantic relatedness to *WordNet*, which is a huge database of English vocabulary. *WordNet* stores the data in a graph structure where synsets or set of synonyms are nodes and relations between them form edges. *WordNet* is lexical database which links the separate databases of English nouns, verbs, adjectives, and adverbs using a hierarchical structure. It is a machine friendly on-line dictionary. The English language consists of forms and senses, where a form is an utterance composed of strings of finite characters and sense is the meaning. A form with a sense is defined as a word in language. In *WordNet* forms are represented by strings and sense is represented by synsets which are sets of synonyms. Each synset is a concept and it represents a node in the hierarchical structure of *WordNet*. The synsets are related to each other by semantic relations which determine the word definitions. Various semantic relations such as Hyponymy(sub-name), Synonymy (symmetric relation), Antonymy(opposing-name), Meronymy(part-name), etc., are defined between words and between synsets or concepts [7]. Hyponymy is a relation between general concept and a specific concept, for example, *Cow - Animal* are hyponyms. Synonymy is a relation between two concepts that have the same meaning. For example, *Beautiful*

- *awesome* are synonyms of each other. Antonymy is a relation between two concepts which are opposite in meaning, for example, *Beautiful - ugly* are antonyms of each other. Meronymy is a relation between a concept which is part of another whole concept, for example, *Hand - finger* are meronyms.

As HSO measure used WordNet graph structure to calculate semantic relatedness between two English words, similarly we use HSO measure with UMLS graph to find semantic relatedness between medical terms.

2.3 Unified Medical Language System

Unified Medical Language System (UMLS) is a large database of biomedical and health data from different available sources. It provides knowledge sources and tools which can be used by developers to build systems for different purposes related to public health or applications in the fields of Bioinformatics and Natural Language Processing.[1] The knowledge sources can be accessed using tools provided by UMLS Technology Services (UTS), or using a JAVA application program called MetamorphoSys.

2.3.1 UMLS Metathesaurus

UMLS Metathesaurus is a large database of biomedical concepts from various multilingual source vocabularies. It contains information about the biomedical concepts such as their definitions, terminologies related to them and the relations between over a million of these concepts. It is built from the huge and diverse medical and health related information available on-line such as different thesauri, statistics, catalogs, biomedical literature, terminologies used in patient care, research information, etc. The Metathesaurus brings together over 100 source vocabularies and helps to relate information contained in them but it is not a vocabulary in itself. Though most of the source vocabularies are in English, Metathesaurus also contains vocabularies in Spanish, French, Dutch, Italian, Japanese, and Portuguese. The information from these source vocabularies is stored in different data, meta data and index files.

UMLS Metathesaurus has over five million terms from the source vocabularies which are grouped into concepts on the basis of their meanings. Each concept has a unique identifier called a Concept Unique Identifier (CUI). Preferred Term is a chosen name for

the concept with multiple names. This Preferred term is computed using ranked source vocabularies.[8]

Concept Unique Identifiers

Concept Unique Identifiers or CUIs are used to identify a unique concept or a meaning and access the information about that concept from the UMLS database. Every time a new concept is added to the Metathesaurus, it is assigned a CUI which is stored in Metathesaurus structure. As the same concept can have different names in different source vocabularies, the UMLS links all the terms which have same meaning or are synonyms of each other in the form of a CUI. All CUIs begin with the letter 'C' which is then followed by seven digits. For example, *Algae* is a Preferred Term for *Alga* and has a CUI *C0002028*. There are also other Unique Identifiers such as Lexical (term) Unique Identifiers (LUI), String Unique Identifiers (SUI), and Atomic Unique Identifiers (AUI).

Sources and Relationships in Metathesaurus

The Metathesaurus consists of different concepts, concept names and other attributes related to different medical terminologies. These concepts are collected from different source vocabularies. Each source vocabulary name is also stored in Metathesaurus tables by a special concept name, 'Intellectual Concept'. Special files in Metathesaurus store the source information along with its version information. These files consist of SAB (Source Abbreviation) as a substring in their names. The Metathesaurus tables store the Root Source Abbreviations in RSAB column and Versioned Source AbbreviationS in VSAB column. Following table shows general information about the sources that are part of Metathesaurus. UMLS Source Documentation has description about all the source vocabularies currently available in UMLS Metathesaurus [9]. Table 2.1 shows the major sources such as SNOMEDCT, MSH, NCBI, Gene Ontology, etc., along with their RSAB.

There are various relations included in the Metathesaurus which relate two concepts within a same source vocabulary by a Intra-source relationship and two concepts from different source vocabularies using Inter-source relationship. Each relationship has a Relation Unique Identifier (RUI). Both synonymous and non-synonymous relationships

Table 2.1: Major sources in UMLS Metathesaurus

Source Name	RSAB
SNOMED Clinical Terms	SNOMED-CT
Mesh	MSH
NCBI Taxonomy	NCBI
Gene Ontology	GO

are maintained in the data tables of Metathesaurus. Each relationship has a label REL which denotes name of that relation. For example, a parent relation (inverse is-a relationship attribute) has a label (REL) as 'PAR'(Parent Relation) and child(is-a relation attribute) relationship has a REL as 'CHD'. These relationships are derived from the hierarchies of the source vocabularies such as SNOMEDCT, MSH, etc., and other medical data. Two concepts can be related to each other by different relations such as is-a or PAR (Parent) relation. For example, referring to Figure 2.3.1 *Entire Anatomical Structure* is related to *Entire Body as a Whole* by PAR relation, whereas *Corpse* is related to *Entire Body as a Whole* by CHD relation. Along with 'PAR' and 'CHD' relationships, others relations such as 'RB'(Broader Relationship), 'RN'(Narrower Relationship), 'RO'(Other Relationship), etc. that symbolize relationships such as association, aggregation, similarity, etc, are found in majority in different source vocabularies.

2.3.2 SNOMEDCT vocabulary

Amongst hundreds of source vocabularies integrated in the UMLS, one of the important and largest source vocabularies is SNOMED Clinical Terms (SNOMEDCT). This is a comprehensive clinical terminology which is used to represent the clinical information [10]. As, SNOMEDCT provides standardized clinical data hierarchies, relations and concepts, it can be used as a reference vocabulary for data analysis by medical researchers such as doctors, software developers and health care organizations [11].

Each of the 311,000 concepts in SNOMEDCT is uniquely identified by a CUI like other source vocabularies in UMLS. It is structured in such a way that each concept is defined using other concepts related to it. A wide variety of relations such as 'PAR', 'CHD', 'RB', 'RN' 'RO', etc. are used to connect the related concepts. The concepts,

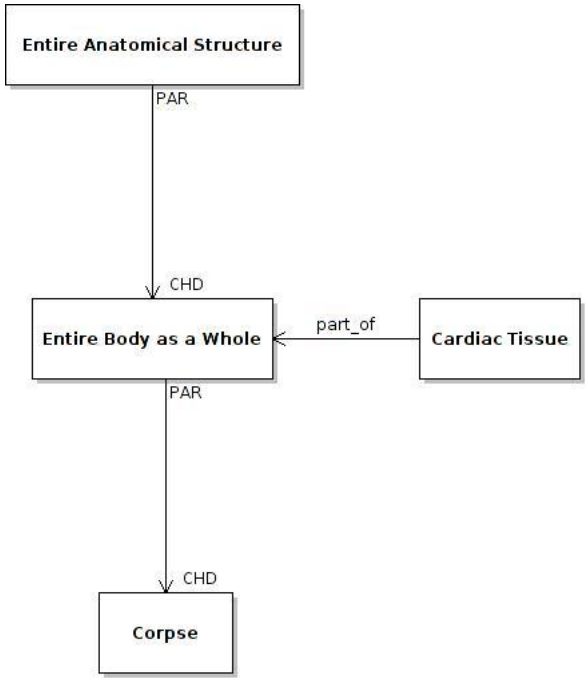

Figure 2.3: Relations in UMLS Metathesaurus

terms and relationships in SNOMEDCT are divided in the different hierarchies such as Procedure, Organism, Substance, Event, etc. SNOMEDCT also contains hundreds of Relation Attributes denoted by RELA which describe the relationships (REL) in detail. For example, *part_of* is a relation attribute and it is represented as follows: *Cardiac Tissue* is part_of *Entire Body as a Whole*, referring Figure 2.3.1. Table 2.2, shows most frequent relation attributes (RELA) from SNOMEDCT along with their relations.

SNOMEDCT vocabulary widely uses 'Defining' relationships, which can used for defining a concept using its relationships with other neighboring concepts [12]. There are around 1,360,000 actively used defining relationships in SNOMEDCT. They are primarily used to provide logical definition of a concept and to perform concept modeling. The defining characteristics of SNOMEDCT concepts consist of 'ISA' relations (PAR/CHD relations) and 'Defining attribute relationships' (Other relations such as RO, RB, RN, etc). Semantic relatedness between two concepts is nothing but, a measure of how closely related two concepts' meanings are. Thus defining attribute relationships

Table 2.2: Most Frequent Relation Attributes in SNOMEDCT

Relation Attribute (RELA)	Relation (REL)	Frequency
isa	CHD	532724
inverse_isa	PAR	532724
same_as	SY	87770
episodicity_of	RO	82243
has_episodicity	RO	82243
has_clinical_course	RO	81791
clinical_course_of	RO	81791
severity_of	RO	81737
has_severity	RO	81737
mapped_to	RN	70084
mapped_from	RB	70084
finding_site_of	RO	69702
has_finding_site	RO	69702
method_of	RO	56233
has_method	RO	56233
has_priority	RO	51337
priority_of	RO	51337
associated_morphology_of	RO	50003
has_associated_morphology	RO	50003
part_of	RN	46647
has_part	RB	46647
has_direct_procedure_site	RO	29948
direct_procedure_site_of	RO	29948
inverse_may_be_a	RO	29610
may_be_a	RO	29610
access_of	RO	28806
has_access	RO	28806
is_interpreted_by	RO	23794
interprets	RO	23794
inverse_was_a	RB	21151
was_a	RN	21151

are a useful way to define the meaning of the concept using its relationships with sibling concepts. Example in Figure 2.3.2 shows how defining relations successfully define the concept. In the following example a term 'Bacterial Pneumonia' can be defined using its neighboring concepts connected by 'ISA' relation and attribute relations 'finding_site' and 'causative_agent'.

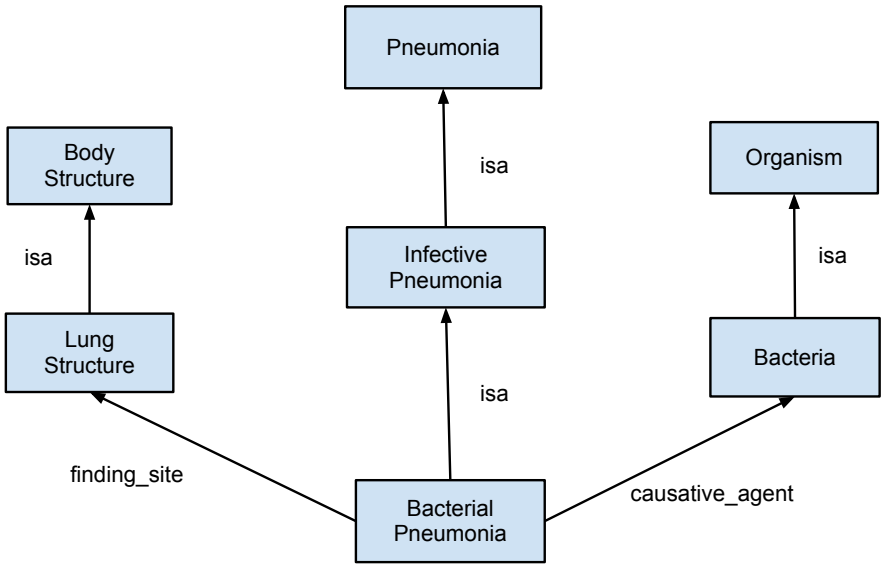

Figure 2.4: Defining attributes used for defining Bacterial Pneumonia

Relation attributes in SNOMEDCT To understand the major defining attribute relationships better, we briefly describe the top relation attributes in SNOMEDCT. Even though the following description list does not cover all attribute relationships from all relations of SNOMEDCT, it covers the top 20 most frequently used defining attribute relationship pairs. Each relationship can be thought of as a complementing pair of relation attributes which describe relation between source concept to destination concept and vice versa. The relation attributes are read from right to left. For example, Cerebral cortex part (C1268167) - [CHD-isa] - (C0228262) Operculum structure, stands for 'Operculum structure' is a child concept of 'Cerebral cortex part'.

Some defining attribute relationships are qualifier relations such as 'severity', 'episodicity', 'priority', 'clinical course', etc. or temporal relations such as 'may_be_a', 'replaced_by', 'was_a', etc. Such relations are not purely based on the meanings of the concept. For example, they might be used used to connect a source concept to its previously used concept name (in case of temporal relationships) or to connect disorders such as 'Fever' and 'Migraine' to 'Severities' to define the severity of the diseases. Thus, these qualifier or temporal relations are beneficial in knowing characteristics of concepts and may be used by UMLS to keep meta-data information about them. Some of the temporal and qualifying relations are also covered in the following list of attributes, to show how these relation attributes are more useful for meta-data purposes than for semantic calculations.

isa and inverse_isa represent Supertype-Subtype relationship between concepts.

Example: (C0228262) Operculum structure - [PAR-inverse_isa] - Cerebral cortex part (C1268167)
Cerebral cortex part (C1268167) - [CHD-isa] - (C0228262) Operculum structure

same_as represents relationship between two concepts which are similar in meaning but differ in spelling or word forms.

Example: (C0332649) Surgical margin involved by tumor - [SY-same_as] - Surgical margins involved by tumour (C0332649)

episodicity_of and has_episodicity represent qualifying relation between a concept

and number of episodes of care provided denoted by a qualifying concept Episodicities.

Example: (C1290036) Disorder of finger - [RO-episodicity_of] - Episodicities (C0565958)
Episodicities (C0565958) - [RO-has_episodicity] - (C1290036) Disorder of finger.

has_clinical_course and clinical_course_of represent qualifying relation between a disease or disorder and its course denoted by qualifier concept Courses (has values such as long, short, etc.)

Example: (C0750729) Courses - [RO-has_clinical_course] - Hearing normal (C0234725)
Hearing normal (C0234725) - [RO-clinical_course_of] - (C0750729) Courses

severity_of and has_severity represent qualifying relationship between a concept and its severity denoted by a qualifier concept Severities(has values such as mild, moderate, severe, etc.)

Example: (C0439793) Severities - [RO-has_severity] - Fungal infection of hair (C0343855)
Fungal infection of hair (C0343855) - [RO-severity_of] - (C0439793) Severities

mapped_to and mapped_from represent relation between a narrower concept and a broader concept to which it is mapped.

Example: (C0273412) Open wound of toes, complicated - [RN-mapped_to] - Open wound of toes with damage to nail (C0451915)
Open wound of toes with damage to nail (C0451915) - [RN-mapped_from] - (C0273412) Open wound of toes, complicated

finding_site_of and has_finding_site represent relation between a disorder and an body structure, where disease is found.

Example: (C1123023) Skin structure - [RO-has_finding_site] - Blister of scalp with infection (C0273622)
Blister of scalp with infection (C0273622) - [RO-finding_site_of] - (C1123023) Skin structure

method_of and has_method represent relation between a procedure and an action taken to complete the procedure.

Example: (C1283169) Monitoring action - [RO-has_method] - Internal fetal monitoring during labor (C0204902)
Internal fetal monitoring during labor (C0204902) - [RO-method_of] - (C1283169) Monitoring - action

has_priority and priority_of represent a qualifier relation between procedure and its priority denoted by a qualifier concept Priorities.

Example: (C0439607) Priorities - [RO-has_priority] - Examination of finger (C0562246)
Examination of finger (C0562246) - [RO-priority_of] - (C0439607) Priorities

associated_morphology_of and has_associated_morphology represent a relation between a disorder/disease and changes in morphological structure caused by the disease.

Example: (C0021368) Inflammation - [RO-has_associated_morphology] - Hip juvenile osteochondropathy (C0410504)
Hip juvenile osteochondropathy (C0410504) - [RO-associated_morphology_of] - (C0021368) Inflammation

part_of and has_part represent a relation between a concept denoting a 'part' and concept denoting a 'whole'.

Example: (C1281591) Entire face - [RN-part_of] - Entire beard (C1280542)
Entire beard (C1280542) - [RB-has_part] - (C1281591) Entire face

has_direct_procedure_site and direct_procedure_site_of procedure and the body structure or site/part which is directly affected by procedure.

Example: (C0040357) Toe structure - [RO-has_direct_procedure_site] - Debridement of toe (C2959629)
Debridement of toe (C2959629) - [RO-direct_procedure_site_of] - C0040357) Toe structure

access_of and has_access represent a relation between a procedure and way used to access a site in the procedure denoted by concept Surgical access values (open, closed, etc.)

Example: (C0920347) Procedure on spinal cord - [RO-access_of] - Surgical access values (C0587266)
Surgical access values (C0587266) - [RO-has_access] - (C0920347) Procedure on spinal cord

is_interpreted_by and interprets represent a relation between an entity or procedure and its interpreted/evaluated form.

Example: (C0314720) Vocal resonance - [RO-interprets] - Egophony (C0231872)
Egophony (C0231872) - [RO-is_interpreted_by] - (C0314720) Vocal resonance

inverse_was_a was_a represent temporal relation between a current concept and its previous concept name.

Example: (C0160323) Heart injury with open wound into thorax, unspecified - [RB-inverse_was_a] - Heart injury, open (C1279560)
Heart injury, open (C1279560) - [RB-was_a] - (C0160323) Heart injury with open wound into thorax, unspecified

has_active_ingredient and active_ingredient_of represent a relation between a drug and its active ingredient.

Example: (C0002645) Amoxicillin - [RO-has_active_ingredient] - Amoxicillin 250mg/ 5mL oral suspension (C0776224)
Amoxicillin 250mg/ 5mL oral suspension (C0776224) - [RO-active_ingredient_of] - (C0002645) Amoxicillin

has_causative_agent and causative_agent_of represent relation between a disorder and organism, substance or physical entity which is a cause for the disease.

Example: (C0275117) Poisoning by honey bee sting - [RO-causative_agent_of] - Honey bee venom (C0440459)
Honey bee venom (C0440459) - [RO-has_causative_agent] - (C0275117) Poisoning by honey bee sting

has_dose_form and dose_form_of represent a relation between a drug product and its dose form.

Example: (C0993159) Oral tablet - [RO-has_dose_form] - Niclosamide 500mg tablet (C0689758)
Niclosamide 500mg tablet (C0689758) - [RO-dose_form_of] - (C0993159) Oral tablet

has_definitional_manifestation and definitional_manifestation_of represent a relation between disorder and the observations which define the disorder

Example: (C0393903) Painful legs and moving toes - [RO-definitional_manifestation_of] - Pain (C0030193)
Pain (C0030193) - [RO-has_definitional_manifestation] - (C0393903) Painful legs and moving toes

uses_device and device_used_by represent a relation between an action and device or instrument used for completing an action in a procedure

Example: (C0392220) Scalpel - [RO-uses_device] - Biopsy of lesion of mesentery of small intestine (C0405843)
Biopsy of lesion of mesentery of small intestine (C0405843) - [RO-device_used_by] - (C0392220) Scalpel

2.3.3 MSH vocabulary

Medical Subject Headings (Mesh/MSH) is another popular vocabulary made available by National Library of Medicine (NLM). MSH can be thought as a hierarchical thesaurus with terms naming descriptors at various levels of the tree. The descriptors are arranged in the hierarchical structure along with alphabetical order and allows searching terms at different levels efficiently. It was first officially published by NLM in 1960 with 4,400 descriptors. It has been continuously developed from then and has expanded to contain 26,581 descriptors (2012 version) till date.

Currently MSH is used by medical libraries for creating catalogs of medical journals and documents. It is also used for indexing and maintaining articles of MEDLINE/PubMED database. Furthermore it is and can be used search engines to retrieve medical documents and search results [13]. Similar to SNOMEDCT, MSH can be accessed through UMLS Metathesaurus web-services along with other access options such

as downloading a copy of database on local machine or using a DVD copy for installation. Additional information about MSH is available on MSH website,[1] where MSH can be obtained in electronic form, without any charge.

MSH consists of hierarchical relationships as fundamental components which are useful for traversing MSH vocabulary. MSH has hierarchical structure similar to a tree structure and currently has 9 levels, where each level represents a degree of specificness. The PAR/CHD pair of relations representing 'is-a' relationship is used by MSH to connect a specific concept to it's general parent concept (similar to SNOMEDCT).

Example: (C0001546) Adjustment Disorders - PAR - Mental Disorders (C0004936) (C0004936) Mental Disorders - CHD - Adjustment Disorders (C0001546)

MSH widely uses 'Sibling' (SIB) relation to represent an associative relation between two related concepts. Being the most widely used association relation (frequency count:426572) in MSH vocabulary, SIB relation can be used to represent horizontal link in HSO's configuration.

Example: (C0034019) Public Health - SIB - Disaster Medicine (C1955980) (C1955980) Disaster Medicine - SIB - Public Health (C0034019)

Along with PAR/CHD and SIB relations MSH consists of relations such as 'RB', 'RN' and 'SY (similar to SNOMEDCT), where 'RB' represents the broader relation, 'RN' represents a narrower relation and 'SY' denotes a synonymous relation to connect terms with different spelling or word form. Relation RB is mostly used with relation attribute 'mapped_from' along with its complementary relation RN with it's relation attribute 'mapped_to'.

Example: (C0243550) pyridosine - [RB-mapped_from] - Amino Acids (C0002520) (C0002520) Amino Acids - [RN-mapped_to] - pyridosine (C0243550)

MSH is a rich source of medical concepts and has wide applications in indexing

[1] http://www.nlm.nih.gov/mesh

and searching the medical databases and cataloging collections. Thus, it is used to perform semantic relatedness experiments and the results are used for comparison with SNOMEDCT results.

2.3.4 Accessing UMLS Data

The UMLS Metathesaurus data was accessible through the UMLS Knowledge Sources (UMLSKS) server until January 2011, after which it was replaced by UMLS UTS (UMLS Terminology Services) server. However, the original UMLSKS server can be accessed by using UMLSKS legacy API[2] along with the newer UTS API 2.0 released in April 2012.[3] We have used the UMLSKS API for accessing UMLS data in both JAVA and PERL implementations, as UTS still provides a backward compatibility towards the old UMLSKS API.

The UMLS Knowledge sources can be accessed using either of the following tools:

- MetamorphoSys: the UMLS installation and Customization Program.
- UMLS Terminology Services (UTS) (formerly UMLSKS)

MetamorphoSys is a Java application which helps you to install and customize the UMLS Knowledge sources locally on the developer's machine. It also helps in making the subset of Metathesaurus based on a particular filters such as a terms within selected source vocabulary or terms with specific Semantic Type. UMLS project also provides a number of free software tools to handle and manipulate UMLS data and they can be obtained through UTS website[4] or though the UMLS DVD.

UMLS Terminology Services: UTS is a set of web-based interactive tools that help developers and users to access UMLS Knowledge sources, UMLS data files and Metathesaurus vocabularies. Developers with valid UMLS Metathesaurus license and UTS account can access the UMLS Knowledge sources over the Internet. Developers can also download all UMLS Release files along with the latest release before they are made available through DVDs. UTS not only provides access to the UMLS data files,

[2] https://uts.nlm.nih.gov
[3] https://uts.nlm.nih.gov//home.html#apidocumentation
[4] https://uts.nlm.nih.gov/home.html

but also allows users to search the UMLS Knowledge Sources such as Metathesaurus and SNOMEDCT using the Metathesaurus browser and SNOMEDCT browser. A user can search about a Metathesaurus concept by entering its name or CUI or a code. Users are provided with different useful search options to limit their search such as to a particular source or release.

Finally, UTS also facilitates developers with UTS Web Services Application User Interface (API) along with the old UMLSKS API. Developers can query the variety of Web Services provided by these APIs and request for information from the Metathesaurus vocabularies and files.[14]

Chapter 3

Algorithm

The aim of this chapter is to explain the details of the overall algorithm used to develop the measure of semantic relatedness for medical terminologies. We first present the general flow of execution and explain each step in detail. We then present a formal discussion of algorithms used for implementing the HSO measure. After the algorithms' details, we provide an insight to the details of both Perl and Java implementations of the algorithm, along with the software packages made available by this work. We also describe the configuration details that help understand the different parameters used in the calculation of semantic relatedness. Please note that both Perl and Java implementations, are based on same algorithms presented in this chapter and are configured in a similar manner. The chapter uses Perl code snippets and examples by default to explain the algorithmic details, whenever necessary.

The measure uses the concept of an allowable path introduced by HSO [6], explained in detail in Section 2.1 and applies it to UMLS Metathesaurus database. The implementation uses UMLSKS Web services to access the huge medical database. As the data is not stored locally on machine, one of the important tasks is to query the UMLS Knowledge Services using different Application Programming Interfaces (APIs) made available by UMLS Technology Servers (UTS)[8]. After accessing the information about the medical terms/CUIs, the next task is to apply the HSO measure to find the Semantic Relatedness between these terms. To use the HSO measure, an allowable shortest path between the input CUIs is calculated by forming a bidirectional graph of the input

CUIs along with the neighboring CUIs. A customized Breadth First Search and Dijkstra's Algorithm is used to find the shortest path between two nodes in the graph. The changes in the directions and the number of nodes in this shortest allowable path lead to calculating the Semantic Relatedness value using the HSO formalization[6].

The overall flow of execution can be divided into following sections :

1. Getting User Details : Get the username and password details of UTS account from user
2. Authenticate the user : Authenticate the user using UMLSKS authentication services and SOAP::Lite[15]
3. Accept the input terms/CUIs : Accept two medical terms/CUIs between which the semantic relatedness will be measured and verify the accuracy of the CUIs
4. Query the Web Service : Query the UMLSKS Web Service with each of the input CUIs to get back the information about the neighbors of the input CUIs
5. Filter and Store the results : Filter the required neighbors' information for the query CUI from the results returned from the Web Service and store this information
6. Form graph : Form a bidirectional graph of the input CUIs connected with their neighbor CUIs using a BFS algorithm.
7. Find Shortest Allowable path : Find the shortest allowable path using Dijkstra's Algorithm and return the path information.
8. Get Semantic Relatedness value : Use HSO formula and shortest path information to find the Semantic Relatedness between the input CUIs.

3.1 Accessing UMLSKS using Web Services

UMLS Knowledge Sources are accessed using the UMLSKS APIs supported by UTS. A user is required to have a valid UTS account and a license to access the UMLS Metathesaurus. A Simple Object Access Protocol (SOAP-Lite) developed for Perl users

is used for Web Service message passing. A SOAP-Lite engine is used to route the Web Service messages to and from the UMLSKS[15] in Perl implementation.

3.1.1 Authenticating User

User's username and password details are accepted and the user is validated using the UMLS authentication server[8]. The authentication process consist of following steps :

1. Initialize UMLS Authentication service.
2. Obtain proxy granting ticket using the user's username and password details.
3. Obtain proxy ticket using the proxy granting ticket.
4. Initialize UMLSKS web service using SOAP::Lite.

3.1.2 Verification of Input

The user enters two terms or CUIs, between which the Semantic Relatedness is to be calculated. If the inputs entered are terms such as *Disease*, *Blood*, etc., then the Web Service named *findCUIByExact*, which accepts a term and returns all CUI(s) for that term by matching the exact string. Thus, CUIs for the input terms are obtained. If the inputs given by the user are CUIs then these CUIs are validated. A valid CUI is a string that starts with letter 'C' followed by seven digits. If the input CUIs are invalid, the user is notified.

3.1.3 Accessing data

The UMLS provides a Web Service API that consists of variety of Web Services which query the UMLS database and return the required results to the user. Some of the Web Services that access the UMLS Metathesaurus data are *findCUIByExact*, *getConceptProperties* and *describeSources*[16].

findCUIByExact: The Web Service *findCUIByExact* locates the CUI for input term by matching the input term exactly[16]. This Web Service accepts various parameters such as

- the *casTicket*, which is the proxy ticket obtained during the authentication process,
- the *searchString*, which is the query term entered by user in the SOAP::Lite type,
- the *language*, vocabulary language to use,
- the *release*, which accepts the UMLS release version to be used for searching,
- the *SABs* (Source Abbreviations) flag, that expects the list of source vocabularies that should be included in the search.

The flags such as *SABs*, *language* and *release* restrict the search scope and return the required specific results. One sample query to *findCUIByExact* is as shown in Figure 3.1. In the query shown in Figure 3.1, the *language* parameter is set to 'ENG'(English), *release* is set to '2010AB' and sources or *SABs* is set to 'SNOMEDCT'.

Figure 3.1: Sample query of findCUIByExact in Perl

```
 my result_refernce = run_query($service,
'findCUIByExact',
                  {
                   casTicket => $proxyticket,
                   searchString => SOAP::Data->type(string => 'Blood'),
                   language => 'ENG',
                   release => '2010AB',
                   SABs => [qw( SNOMEDCT )],
                  },
                 );
```

All the Web Services related to UMLS Metathesaurus return the results as a hash reference in the JASON format, which is hash of list of hash recursively. The sample output of the query made in Figure 3.1 is shown in Figure 3.1.3.

getConceptProperties: This Web Service accepts a CUI and returns various properties of the CUI. The properties include the details of all the CUIs related to the input CUI, definitions, terminologies and semantic designations of the concept[16]. The user can specify different flags which decide the results returned Web Service. Some of the important flags that can be included are :

contentClassName	gov.nih.nlm.kss.models.meta.concept.ConceptId
performanceMode	0
contents	CN: Blood performanceMode: 0 CUI: C0005767 key: ** undefined **
empty	0
release	2010AB
queryInput	casTicket: ST-138343-0PbMhuqKZ7LxbucJKtTc-cas language: ENG includeSuppressibles: 0 release: 2010AB CVF: 0 SABs: SNOMEDCT searchString: Blood
key	** undefined **

Figure 3.2: Sample output of findCUIByExact

- *SABs*, which accepts list of source vocabularies that should be included in searching,
- the *casTicket* which is the proxy ticket obtained during the authentication process,
- the *CUI*, which is the query CUI,
- the *language*,
- the *release*, which accepts the UMLS release version to be used for searching,
- *includeRelations*, which should be set to true when relationships' details of query CUI are expected in results,
- *relationTypes*, which accepts list of relation types,
- *includeDefinitions*, which is set to tru when definitions of the query CUI are needed,

Figure 3.3 shows sample query made to Web Service *getConceptProperties*, with the query *CUI* being the CUI for term 'Blood', the *relationTypes* set to 'PAR'(parent)

relation. The definitions are not included in the query, *SABs* flag is set to 'SNOMEDCT' source vocabulary, *language* is set to 'ENG'(English) and UMLS *release* 2010AB is used.

Figure 3.3: Sample query of getConceptProperties in Perl

```
my $return_ref = run_query($service,
'getConceptProperties',
                      {
                        casTicket => $proxyticket,
                        CUI => SOAP::Data->type( string => 'C0005767' ),
                        language => 'ENG',
                        release  => '2010AB',
                        SABs => [qw( SNOMEDCT )],
                        includeDefinitions   => 'false',
                        includeRelations => 'true',
                        relationTypes    => ['PAR'],
                },
        );
```

3.2 Determining value of constants C and k in HSO formulation

The HSO formulation not only considers the length of the path joining the concepts, but also takes into account the direction of the path. It penalizes a path if it has more changes in direction and discardS a path if it does not fit into the allowable patterns set shown in Figure 2.1.2. Each allowable pattern is formed by interconnected vectors and it is allowed for a vector to extend without restriction in either (up, down and horizontal) direction.

The HSO formulation to calculate the semantic relatedness value is:

$$Semantic\_relatedness = C - path\_length - k * changes\_in\_direction$$

Here the values of C and k are not predefined by HSO. The values were determined by experimenting with different possible values and making valid assumptions.

- Observation 1: It was observed that the maximum length of the path in results obtained for experimental data sets using up and down links, was 14.
- Assumption 1 : Thus the maximum length of the path was assumed to be $<= 15$. Even if the path length > 15 is possible, it will not yield to meaningful relatedness, as the longer the path, the less are the two concepts related.
- Observation 2 : The maximum number of direction changes in any allowable path as suggested by HSO is 2.
- Assumption 2 : Thus it was assumed that the maximum penalty for number of changes in direction would be $k * 2$.
- Observation 3 : The maximum value of $path\_length - k * changes\_in\_direction$, would be definitely less than the assumed $path\_length = 15$.
- Assumption 3 : The value of C was assumed to be equal to 20 so that the semantic relatedness value does not fall below 0.
- Observation 4 : As the HSO formulation allows only one direction change (with an exception of change made with horizontal link), the penalty should increase considerably with each direction change, i.e., the penalty should not simply be linear.

Experiments were done with different values of k taking into consideration the fourth observation, but as shown in Table 3.1, all the absolute values of k had similar effect on the penalty, as it did not take into account the respective path length between the CUI pairs. Thus, the chosen values of k did not affect Spearman's correlation values with the human judgments and resulted in same correlation values. Initially the value of k was chosen to be $1/4$. But, these experiments with different values of k led to another observation : 'The value of k should be dependent on the path length'. Thus value of k was multiplied by the initial relatedness value calculated as :

$$semantic\_relatedness = initial\_relatedness - (k * initial\_relatedness) * change\_in\_direction$$

where,

$$initial\_relatedness = C - path\_length$$

where C = 20. Thus the penalty was applied in proportion to the path length.

With the values of C and k substituted with their determined values, the equations to calculate the semantic relatedness are:

$$initial\_relatedness = 20 - path\_length$$

$$semantic\_relatedness = initial\_relatedness - (initial\_relatedness/4) * change\_in\_direction$$

To verify that the value of relatedness can never be negative using this formulation, consider a corner case where $path\_length = 15$ and $changes\_in\_direction = 2$, the value of semantic relatedness would be 2.5.

$$initial\_relatedness = 20 - 15 = 5$$
$$semantic\_relatedness = 5 - (5/4) * 2 = 2.5$$

Thus for the concepts sharing extra strong relation, $path\_length = 0$ and $changes\_in\_direction = 0$, thus the semantic relatedness is equal to C, i.e. 20.

Table 3.1: MiniMayoSRS Correlation Values for different values of constant k

k	MiniMayo.coders	MiniMayo.physicians
1/2	0.5390	0.3723
1/3	0.5390	0.3723
1/4	0.5390	0.3723
1/5	0.5390	0.3723

3.3 Calculating Semantic Relatedness

UMLS Metathesaurus is a big graph of medical terminologies, where CUIs form the nodes of the graph and these CUIs are connected to each other with various relations

such as 'PAR', 'CHD', 'RB', 'RN', etc[1]. The Web Service *getConceptProperties* can be used to obtain the CUIs that are related to the input CUIs with different relations. This in memory sub graph is formed by 'FormGraph' algorithm and is passed to the 'FindShortestAllowablePath' algorithm, which finds the shortest available path between the source input term and destination input term, if such path exists using the sub graph. Once a shortest allowable path is obtained, Semantic Relatedness is calculated using the HSO formalization[6].

3.3.1 Graph Formation

The sub graph of UMLS CUIs is formed by 'FormGraph' algorithm which is inspired from the standard Breadth First Graph Traversal algorithm. To avoid the multiple and unnecessary requests to the Web Service, each neighbor CUI is visited only once and a local sub graph is updated every time a CUI is visited. Each node in the subgraph is stored with the least cost of reaching the node from the input node and the path with which it can be reached. Only the nodes that can be reached with an allowable path from source or destination are chosen to expand the graph by querying for their neighbors from UMLS. There is no need to bring the neighbors for the nodes that cannot be reached with an allowable path as they would never be part of an allowable path between source and destination. Every time a subgraph of input CUIs along with useful neighboring CUIs is updated, it is passed to GetAllowableShortestPath algorithm to get back the shortest allowable path between source and destination if it exists. When the first allowable shortest path is found between the source and destination, it is stored as the current available shortest path and its information such as cost, direction changes and direction vector are remembered. The subgraph is expanded by adding new CUI nodes and all the nodes that can be reached with cost less than the cost of current available shortest path are visited. Those nodes which can be reached with cost equal or greater than the cost of current available cost are ignored as they will surely lead to path greater than current shortest path. This ensures that the shortest path obtained at the end of algorithm is the shortest possible path. The detailed algorithm 'FormGraph' is explained in Algorithm 1.1.

Algorithm 3.3.1 FormGraph

Require: $source \neq empty$, $destination \neq empty$, source is valid, destination is valid

$priority\_queue \leftarrow (source, destination)$
$nodeHash \leftarrow empty, subgraphHash \leftarrow empty$
$constK \leftarrow 1/4, constC \leftarrow 20$
$currentAvailableCost \leftarrow MAX, currentShortestPath \leftarrow empty$
$changesInDirection \leftarrow -1, shortestPathDirection \leftarrow empty$
$shortestpathInfo \leftarrow empty$
$neighbors \leftarrow empty$
$nodeHash\{source\} \leftarrow 0, nodeHash\{destination\} \leftarrow 0$
while $priority\_queue \neq empty$ **do**
  $currentNode \leftarrow pop(priority\_queue)$
  $costUptoNode \leftarrow nodeHash\{currentNode\}$
  $currentNode$ is $Visited$
  if $pathUptoNodeisnotallowed$ **then**
    $next$
    {/* Ignore this node if it cannot be reached with an allowable path */}
  end if
  if $costUptoNode >= currentAvailableCost$ **then**
    $next$
    {/* Ignore this node as this may lead to longer path than currentAvailablePath */}
  end if
  $neighbors \leftarrow getNeighbors(currentNode)$
  if $neighbors = empty$ **then**
    $next$
    {/* No neighbors information for this node in UMLS */}
  else $\{neighbors \neq empty\}$
    for all N in $neighbors$ **do**
      if N is $Parent$ **then**
        $subgraphHash\{N\}\{currentNode\} \leftarrow D$
        $subgraphHash\{currentNode\}\{N\} \leftarrow U$

else {N is $Sibling$}
$subgraphHash\{N\}\{currentNode\} \leftarrow H$
$subgraphHash\{currentNode\}\{N\} \leftarrow H$
end if
{/*Here, upward link = U, downward link = D and horizontal link = H */}
if N is not $Visited$ **then**
if N is $Parent$ **then**
$costUptoNeighbor \leftarrow costUptoNode + parentCost$
else {N is $Sibling$}
$costUptoNeighbor \leftarrow costUptoNode + siblingCost$
end if
Store the path direction vector to reach N
if N is in $nodeHash$ **then**
Insert new $costUptoNeighbor$ only if it is smaller than existing cost in $nodeHash$
else {N not in $nodeHash$}
Insert N in $nodeHash$ with $costUptoNeighbor$
$push(priority\_queue, N)$
end if
end if
end for
end if
$shortestPathInfo \leftarrow getShortestAllowablePath(source, destination, subgraphHash)$

if $shortestPathInfo$ is not empty **then**
$currentShortestPath \leftarrow pop(shortestPathInfo)$
$currentAvailableCost \leftarrow pop(shortestPathInfo)$
$changeInDirection \leftarrow pop(shortestPathInfo)$
$pathDirection \leftarrow pop(shortestPathInfo)$
else {$shortestPath$ does not exist}
$next$
{/* Continue with the while loop as currently there is no allowable path between

source and destination */}
end if
end while
if *currentShortestPath* is not empty **then**
initialRelatedness = *constC* − (*currentAvailableCost*/10)
if *changeInDirection* == −1 **then**
changeInDirection ← 0
end if
semanticRelatedness ← *initialRelatedness* − ((*constK* ∗ *initialRelatedness*) ∗ *changeInDirection*)
return *semanticRelatedness*
else {*currentShortestPath* is empty}
return No path exists between *source* and *destination*
end if

3.3.2 Finding Shortest Allowable path

To get the allowable shortest path between the input CUIs, Dijkstra's Algorithm is implemented using the subgraph formed by the *FormGraph*. The standard Dijkstra's Algorithm based on BFS is used with slight variation to get an allowable shortest path instead of any shortest path. Every time an intermediate node is visited by the Dijkstra, the intermediate path connecting the source CUI and the neighbor is checked against an allowable path regular expression which defines HSO's set of allowable patterns[6] to check if this partial path is allowed. If the partial path is not allowed then there is no need to go further in the direction of current node. Thus the node is ignored and Dijkstra continues by visiting next neighbor. If the partial path is allowed, then the cost of the partial path is calculated. The cost of vertical paths/links is less that the cost of horizontal paths, because, the concepts related to the source concept with horizontal relations such as association, generalization, etc., digress more from the meaning of the source concept. We check if the current node can be reached with less cost than the current path, if so, then this partial path is ignored. After all the nodes of the graph are traversed, shortest allowable path is returned if found, else -1 is returned. The algorithm is described in detail in Algorithm 1.2.

Algorithm 3.3.2 GetAllowableShortestPath

Require: *subgraph* from FormGraph, $source \neq empty, destination \neq empty, REGEX$

$shortestpathInfo \leftarrow empty$

$parentCost \leftarrow 1, siblingCost \leftarrow 2$

$allowablePatternRegex \leftarrow REGEX$

$shortestPathCost \leftarrow MAX$

$changesInDirection \leftarrow -1$,$shortestPathDirection \leftarrow empty$

for all $Node$ in *subgraph* **do**

/* Initialize the distance in which this node can be reached from source to infinity */

if $Node$ is *source* **then**

$Node\_distance \leftarrow 0$

else

```
    Node_distance ← infinity
    Mark Node unvisited
  end if
  push(priority_queue, Node)
end for
while priority_queue ≠ empty do
  /* While priority_queue is not empty traverse the subgraph */
  currentNode ← pop(priority_queue)
  if Node_distance = infinity then
    /* No node can be reached from here so stop the search by breaking from while
    */
    Break
  end if
  for all currentNode in neighbors of poppedNode do
    /* Traverse all the neighbors of the popped node in the priority_queue using
    subgraph */
    if currentNode is visited then
      /* Ignore this currentNode */
    else {currentNode is not visited}
      Calculate the new distance and path for currentNode
      Calculate the currentNode_pathDirection using the updated currentNode_path

      /* Check if the currentNode_pathDirection is allowed or not, if not allowed
      then ignore the node */
      if currentNode_pathDirection =~ allowablePatternRegex then
        /* If pathDirection is allowed update the distance if shorter than previous
        distance and direction of currentNode and adjust it's postion in queue*/
        push(priority_queue, currentNode with updated distance)
      end if
    end if
  end for
  Mark currentNode as visited
```

if $currentNode = destination$ **then**
/* Shortest path found between source and destination, save the path information and break from while loop */
$shortestPath \leftarrow currentNode\_path$
$shortestPathDirection \leftarrow currentNode\_direction$
$shortestPathCost \leftarrow getCost(currentNode\_path)$
Break
end if
end while
if $shortestPath \neq empty$ **then**
Calculate number of changesInDirection for the shortest path
$push(shortestpathInfo, shortestPath)$
$push(shortestpathInfo, shortestPathCost)$
$push(shortestpathInfo, changesInDirection)$
$push(shortestpathInfo, shortestPathDirection)$
return $shortestpathInfo$
else
return -1
end if

Algorithm 3.3.3 getCost(aPath)

$currentPathCost \leftarrow 0, direction \leftarrow empty$
for $0 \leq index \leq length(aPath) - 1$ **do**
  $firstNode \leftarrow aPath[index]$
  $secondNode \leftarrow aPath[index + 1]$
  $direction \leftarrow direction$ between $firstNode$ and $secondNode$ from $subgraphHash$
  if $direction = up \mid down$ **then**
    /* Here the cost of upward link (U) and downward link (D) is equal to the $parentCost$*/
    $currentPathCost \leftarrow currentPathCost + parentCost$
  else $\{direction = horizontal\}$
    /* Here the cost of horizontal link (H) is equal to the $siblingCost$*/
    $currentPathCost \leftarrow currentPathCost + siblingCost$
  end if
end for
return $currentPathCost$

3.4 Implementation and Configuration Details

Initially the algorithm was implemented using the Perl programming language. This implementation is made freely available through a CPAN software package known as WebService::UMLSKS::Similarity. This package includes a collection of tools/programs that can be used independently or together as a measure of semantic relatedness in any software application.[1] Webservice::UMLSKS::Similarity uses SOAP-Lite to access UMLS web services and can be executed with Perl version 1.5 and higher, on smaller data scale. As UMLS database is accessed over the network, it became difficult to handle large amount of data (more than million CUIs) efficiently with the Perl implementation. Thus, a Java implementation was used to perform experiments on larger scale. The Java implementation known as Webservice::UMLSKS::Similarity::Java is made available as a Sourceforge open source project.[2]

Webservice::UMLSKS::Similarity::Java uses Java's strengths such as efficient caching, multithreaded prefetching of data from UMLS, efficient memory management and thus allows experiments with larger data sets. It also takes advantage of the Java API provided by UMLS's UTS. Both the Perl and Java implementations can be configured with similar configuration files and are provided with detailed documentation using Perldoc and Javadoc respectively.

Configuration details: The HSO allowable path set shown in Figure 2.1.2 consists of connected vector patterns in three directions, up (U), down (D) and horizontal (H). Relations such as PAR (parent) can be thought as U link, whereas CHD (child) relation can be thought as D link. Other relations such as RB (broader), RN (narrower), RO (others), SY (similar, concepts with different spellings) can be considered as H links.

A Configuration file is used to set the various configurations such the source name (SAB), relations (REL) and relation attributes (RELA). The file is also used to specify what direction should be associated with a particular relation. For example, to specify that the PAR relation should be treated as a U link while calculating the allowable path patterns, the options REL and DIR are used in sample configuration file shown in Figure 3.4. The allowable path patterns set can also be supplied to the program as a

[1] http://search.cpan.org/dist/WebService-UMLSKS-Similarity/
[2] https://sourceforge.net/p/umlsks-sr-java

configuration parameter. It is accepted through a patterns file which consists a regular expression formed by using symbol 'U' for upward link, 'D' for downward link and 'H' for horizontal link. For example, an allowable path pattern that consists of a vector in upward direction followed by a vector in downward direction can be represented as a regular expression `/\bU+D+\b/,`which denotes that any path that consists of an upward vector followed by a downward vector is allowed. Such pattern is similar to the second pattern in the first row in Figure 2.1.2. If the patterns are not supplied as the command line parameter, default patterns are used by the program. The default allowable pattern regular expression that maps the original HSO allowable pattern set is shown in Figure 3.5.

SAB :: include SNOMEDCT
REL :: include PAR,RB
DIR :: include U,H
RELA :: include RB-has_part

Figure 3.4: Sample configuration file

```
/\bU+\b|\bU+D+\b|\bU+H+\b|\bU+H+D+\b|\bD+\b|
\bD+H+\b|\bH+D+\b|\bH+\b/
```

Figure 3.5: Default HSO allowable patterns regular expression

By default the packages are configured to use UMLS release 2012AA and the cost of each upward and downward link is equal to 1 and the cost of one horizontal link is 2. This difference in cost is just a way of making sure that the horizontal paths are expensive over up and down paths, as per HSO's suggestion.

Chapter 4

Experimental Data

The correctness of semantic relatedness measures can be evaluated either by directly comparing their results with human judgments or by measuring the performance of applications using these measures. Rubenstein and Goodenough [17] used direct evaluations against human judgments to evaluate their efforts of finding semantic similarity, whereas some measures of word sense disambiguation [18] and malapropism detection [6] have used these applications to evaluate the performance of the measure. To examine and evaluate the measure developed for semantic relatedness using HSO formulation, different semantic relatedness gold standards are chosen. The evaluation is done by directly comparing the results with the expert's judgments using Spearman's correlation coefficient. These gold standards are made available by the University of Minnesota Pharmacy Informatics lab [19] for researchers to quantify the performance of their measure. The gold standards consist of the set of concept pairs i.e. CUI pairs, which are manually rated and assigned a semantic relatedness value. In this chapter, we present the experimental data with its gold standards and explain how it was formed. We then present the 'big subsets' formed during this thesis work along with their gold standards. Finally we explain the Spearman's correlation coefficient which is used for evaluation of the HSO measure when tested against the gold standards.

4.1 Rubenstein and Goodenough's Data

Rubenstein and Goodenough [17] created pairs of manually rated concepts for English terms for direct evaluation of the relation between the similarity of context and synonymy. In 1965, Rubenstein and Goodenough created a test data of 65 word pairs and obtained synonymy judgments from 51 human subjects. The raters rated the terms on scale of 0.0 to 4.0, where the score of 0.0 meant 'semantically unrelated' and 4.0 meant 'highly synonymous'. Miller and Charles [1991] created a smaller data set of 30 word pairs out of the 65 word pairs of Rubenstein and Goodenough's data. They chose 10 word pairs each from 3 levels of similarity of data, i.e., *high level set (between 4.0 - 3.0)*, *intermediate level set (between 3.0 - 1.0)* and *lower level set (between 1.0 - 0.0)*. The data was rated by 38 subjects on their similarity judgments on the same scale of 0.0 to 4.0.

4.2 MayoSRS and MiniMayoSRS data sets

To evaluate the semantic relatedness between terms in the biomedical domain, a physician at the Mayo Clinic trained in Medical Informatics generated a set of 120 term pairs by following the methodology of Rubenstein and Goodenough. The data set consisted of 30 term pairs from each of the categories from not related at all (1.0) - very closely related (4.0). The term pairs were annotated by 13 medical coders on the wider scale of 1-10, which was narrowed down to match the Rubenstein and Goodenough's scale of 0.0 - 4.0.

A reduced MayoSRS set consisting of 101 medical concept pairs was created from these 120 pairs. The semantic relatedness value for this set is manually assigned by experienced medical coders to form the MayoSRS.gold standard. Though these coders were not formally educated in medicine, they were considered to be good candidates for the task of annotation as they were highly experienced with the use of medical terminologies.

To derive a more reliable test set Pedersen et al. (2009), created the MiniMayoSRS set of 30 pairs. This subset was then annotated by three physicians specialized in rheumatology and 9 medical coders, to form two gold standards, MiniMayoSRS.physicians and MiniMayoSRS.coders respectively. Each pair was annotated on a 4 point scale :

practically synonymous (4.0), related (3.0), marginally related (2.0) and unrelated (1.0). One of the term pairs (lung inltrates) was excluded from this test bed as it was absent in SNOMEDCT (one of the sources in UMLS). Thus, Pedersen et al. [20] created a test set of 29 medical concept pairs that were scored by human experts according to their relatedness.

The format of the test file is :

CUI1 <>CUI2

For example, following is one of the CUI pairs from MiniMayoSRS.cuis set :

C0156543 <>C0000786

Thus, all these test sets have list of concept pairs and the key set for these test sets is of the form :

SemanticRelatedness <>CUI1 <>CUI2

For example, following is the semantic relatedness value for sample line :

3.3 <>C0156543 <>C0000786

Table 4.1 shows the Medical Coders High Reliability Subset or MiniMayoSRS consisting of 29 concept pairs. The gold standards MiniMayoSRS.physicians and MiniMayoSRS.coders are also tabulated which represent the semantic relatedness values assigned to MiniMayoSRS CUI pairs by physicians and coders respectively. The MayoSRS test set is also tabulated along with the MayoSRS.gold standard's semantic relatedness values for the 101 CUI pairs in table A.1, table A.2, and table A.3.

Table 4.1: MiniMayoSRS test set

Term 1(CUI 1)	Term 2(CUI 2)	Coder	Physician
Renal failure(C0035078)	Kidney failure(C0035078)	4.0	4.0
Abortion(C0156543)	Miscarriage(C0000786)	3.3	3.0
Heart(C0018787)	Myocardium(C0027061)	3.0	3.3
Metastasis(C0027627)	Adenocarcinoma(C0001418)	1.8	2.7
Pulmonary brosis(C0034069)	Lung cancer(C0242379)	1.4	1.7
Brain tumor(C0006118)	Intracranial hemorrhage(C0151699)	1.3	2.0
Rheumatoid arthritis(C0003873)	Lupus(C0409974)	1.1	2.0
Pulmonary embolus(C0034065)	Myocardial infarction(C0027051)	1.2	1.7
Antibiotic(C0003232)	Allergy(C0020517)	1.2	1.7
Depression(C0011581)	Cellulitis(C0007642)	1.0	1.0
Diarrhea(C0011991)	Stomach cramps(C0344375)	1.3	2.3
Multiple sclerosis(C0026769)	Psychosis(C0033975)	1.0	1.0
Mitral stenosis(C0026269)	Atrial brillation(C0004238)	1.3	2.3
Congestive heart failure(C0018802)	Pulmonary edema(C0034063)	1.4	3.0
Lymphoid hyperplasia(C0333997)	Laryngeal cancer(C0007107)	1.0	1.3
Diabetes mellitus(C0011849)	Hypertension(C0020538)	1.0	2.0
Carpal tunnel syndrome(C0007286)	Osteoarthritis(C0029408)	1.1	2.0
Xerostomia(C0043352)	Alcoholic cirrhosis(C0023891)	1.0	1.0
Peptic ulcer disease(C0030920)	Myopia(C0027092)	1.0	1.0
Appendicitis(C0003615)	Osteoporosis(C0029456)	1.0	1.0
Hyperlipidemia(C0020473)	Metastasis(C0027627)	1.0	1.0
Cortisone(C0010137)	Total knee replacement(C0086511)	1.0	1.7
Acne(C0702166)	Syringe(C0039142)	1.0	2.0
Stroke(C0038454)	Infarct(C0021308)	2.8	3.0
Varicose vein(C0042345)	Entire knee meniscus(C0224701)	1.0	1.0
Rectal polyp(C0034887)	Aorta(C0003483)	1.0	1.0
Delusion(C0011253)	Schizophrenia(C0036341)	2.2	3.0
Cholangiocarcinoma(C0206698)	Colonoscopy(C0009378)	1.0	1.3
Calcication(C0175895)	Stenosis(C0009814)	2.0	2.7

4.3 UMNSRS_reduced_rel and UMNSRS_reduced_sim test sets

The UMNSRS_reduced_rel and UMNSRS_reduced_sim test sets are created at the University of Minnesota Pharamacy Informatics Lab for quantifying semantic relatedness and semantic similarity respectively between medicals term pairs from the UMLS [21]. The UMNSRS_reduced_rel set consists of 430 CUI pairs and UMNSRS_reduced_rel.gold standard contains the semantic relatedness values assigned to them. UMNSRS_reduced_sim consists of 401 CUI pairs along with UMNSRS_reduced_sim.gold standard which contains semantic similarity values assigned to the CUI pairs.

The initial test sets were created by choosing concepts from the UMLS with one of these semantic types: disorders, symptoms and drugs. A practicing physician manually selected 30 term pairs with at least one single-word term. The semantic relatedness categories such as completely unrelated, somewhat unrelated, somewhat related, and closely related are covered in these test sets by selecting 30 term pairs for each of them and in 6 relation type categories such as : DISORDER-DISORDER, DISORDER-SYMPTOM, DISORDER-DRUG, SYMPTOM-SYMPTOM, SYMPTOM-DRUG, DRUG-DRUG. Thus an initial set of 724 samples was developed for the study which covers variety of medical terminologies. The process of annotating these term pairs was conducted by 8 medical residents at the University of Minnesota Medical school participated in the study.

The medical experts gave a semantic relatedness and semantic similarity value to each pair in the test set based on their intuition. On the relatedness task, all raters succeeded on 587 (81%) of 724 samples and on the similarity task, 566 (78%) of 724 pairs were successfully completed by all. Thus from these samples, two reduced sets UMNSRS_reduced_rel (430 samples) and UMNSRS_reduced_sim (401 samples) were formed taking into consideration the agreements and dis-agreements of the raters [21] along with two keys, viz., UMNSRS_reduced_rel.gold and UMNSRS_reduced_sim.gold respectively which can be used as gold standards.

4.4 MiniMayoSRS set for MSH vocabulary

The original MiniMayoSRS test set was formed by using the concepts from the SNOMECT source vocabulary of the UMLS. The University of Minnesota Pharmacy Informatics lab [19] has also made available a gold standard for the CUI pairs of MiniMayoSRS test set that occur in MSH source vocabulary. The subsets for MSH are formed by including all the CUI pairs from MiniMayoSRS test set, that are present in MSH vocabulary. Table 4.2 shows the MiniMayoSRS.msh.coders and MiniMayoSRS.msh.physicians gold standards respectively. Similar to the MiniMayoSRS.coders and MiniMayoSRS.physicians gold standards, these standards are also scored on the scale of 0.0 - 4.0.

Table 4.2: MiniMayoSRS.msh test set

(CUI 1)	(CUI 2)	Coder	Physician
Failure, Kidney(C0035078)	Failure, Kidney(C0035078)	4.0	4.0
Heart, NOS(C0018787)	Myocardium, NOS(C0027061)	3.0	3.3
Cerebrovascular accident, NOS(C0038454)	Infarction, NOS(C0021308)	2.8	3.0
Legal abortion procedure(C0000812)	Abortions, spontaneous(C0000786)	3.3	3.0
Delusion, NOS(C0011253)	Schizophrenia NOS(C0036341)	2.2	3.0
Heart failures(C0018801)	Oedema - pulmonary NOS(C0034063)	1.4	3.0
Metastasis, Neoplasm(C0027627)	Adenocarcinoma, NOS(C0001418)	1.8	2.7
Vehicles, Motor(C0175845)	Stenose(C1261287)	2.0	2.7
Stenose(C1261287)	Fibrillation, Atrial(C0004238)	1.3	2.3
Rheumatoid arthritis NOS(C0003873)	Lupus Vulgaris(C0024131)	1.1	2.0
Neoplasms, Brain(C0006118)	Hemorrhages, Intracranial(C0151699)	1.3	2.0
Syndrome, Carpal Tunnel(C0007286)	Degenerative polyarthritis, NOS(C0029408)	1.1	2.0
Diabetes mellitus NOS(C0011849)	Hypertensive disease NOS(C0020538)	1.0	2.0
Acne, vulgaris(C0001144)	Syringe, NOS(C0039142)	1.0	2.0
Agents, Anti-Bacterial(C0279516)	Hypersensitivity NOS(C0020517)	1.2	1.7
Cortisones(C0010137)	Knee Replacement Arthroplasty procedure(C0086511)	1.0	1.7
Fibrosis, Pulmonary(C0034069)	Neoplasms, Lung(C0024121)	1.4	1.7
Cholangiocarcinomas(C0206698)	Colonoscopy, NOS(C0009378)	1.0	1.3
Hyperplasia, NOS(C0020507)	Neoplasms, Laryngeal(C0023055)	1.0	1.3
Depression, Mental(C0011570)	Cellulitis, NOS(C0007642)	1.0	1.0
Multiples sclerosis(C0026769)	Disorders, Psychotic(C0033975)	1.0	1.0
Xerostomia(C0043352)	of alcoholic liver cirrhosis(C0023891)	1.0	1.0
Ulcer, Peptic(C0030920)	Myopias(C0027092)	1.0	1.0
Appendicitis NOS(C0003615)	Osteoporosis NOS(C0029456)	1.0	1.0
Hyperlipidaemias(C0020473)	Metastasis, Neoplasm(C0027627)	1.0	1.0

4.5 Big subsets for SNOMECT

While performing the experiments with PAR/CHD relations as upward/downward relations and a large number of relations and relation attributes as horizontal relations, the algorithm was not able to find paths for some CUI pairs (around 15%) from the test sets (MiniMayoSRS, MayoSRS, UMNSRS_reduced_rel and UMNSRS_reduced_sim), even after exploring greater than 50,000 neighboring CUIs. As these CUIs have a large number of siblings (approximately 1000) on average, it is difficult to handle and compute relatedness value for such problem CUI pairs. For example, the degree of concept 'Parenteral dosage form product' is 2060, which explodes the data structures used by the program, increasing the time and space complexity of the program. As concepts' information is accessed from UMLS at each level, the local graph structure of concepts increases exponentially.

To experiment with large set of relations and attributes in feasible time, a subset is formed for each test set, excluding the problem CUI pairs. These subsets consist of CUI pairs which have manageable neighbor concepts. Combining CUI pairs from all subsets, the algorithm was able to find correct semantic relatedness value for 814 CUI pairs out of total 961 CUI pairs from original data sets. The individual division of CUIs in the subsets is as follows - MiniMayoSRS test set : 23/29, MayoSRS test set : 84/101, UMNSRS_reduced_rel test set : 366/430, UMNSRS_reduced_sim test set : 341/401. These subsets are termed as 'big subsets' as their size is very close to actual test sets' size.

Table 4.3 and Tables A.8 and A.9 show the big subsets and their keys for MiniMayoSRS and MayoSRS test sets respectively. Similarly subsets are formed for UMNSRS_reduced_rel and UMNSRS_reduced_sim, thus forming a subset pool consisting of total 814 CUI pairs.

4.6 Spearman's Rank Correlation Coefficient Evaluation

Spearman's correlation coefficient is used for evaluation of HSO measure for experimental data sets. The Spearman's rank correlation coefficient (ρ)[1] ,measures the degree to

[1] sometimes known as Spearman's rho

Table 4.3: MiniMayoSRS Big subset test set with key

Term 1(CUI 1)	Term 2(CUI 2)	Coder	Physician
Renal failure(C0035078)	Kidney failure(C0035078)	4.0	4.0
Abortion(C0156543)	Miscarriage(C0000786)	3.3	3.0
Heart(C0018787)	Myocardium(C0027061)	3.0	3.3
Pulmonary brosis(C0034069)	Lung cancer(C0242379)	1.4	1.7
Brain tumor(C0006118)	Intracranial hemorrhage(C0151699)	1.3	2.0
Rheumatoid arthritis(C0003873)	Lupus(C0409974)	1.1	2.0
Pulmonary embolus(C0034065)	Myocardial infarction(C0027051)	1.2	1.7
Antibiotic(C0003232)	Allergy(C0020517)	1.2	1.7
Depression(C0011581)	Cellulitis(C0007642)	1.0	1.0
Diarrhea(C0011991)	Stomach cramps(C0344375)	1.3	2.3
Multiple sclerosis(C0026769)	Psychosis(C0033975)	1.0	1.0
Mitral stenosis(C0026269)	Atrial brillation(C0004238)	1.3	2.3
Congestive heart failure(C0018802)	Pulmonary edema(C0034063)	1.4	3.0
Lymphoid hyperplasia(C0333997)	Laryngeal cancer(C0007107)	1.0	1.3
Diabetes mellitus(C0011849)	Hypertension(C0020538)	1.0	2.0
Carpal tunnel syndrome(C0007286)	Osteoarthritis(C0029408)	1.1	2.0
Xerostomia(C0043352)	Alcoholic cirrhosis(C0023891)	1.0	1.0
Peptic ulcer disease(C0030920)	Myopia(C0027092)	1.0	1.0
Appendicitis(C0003615)	Osteoporosis(C0029456)	1.0	1.0
Cortisone(C0010137)	Total knee replacement(C0086511)	1.0	1.7
Acne(C0702166)	Syringe(C0039142)	1.0	2.0
Stroke(C0038454)	Infarct(C0021308)	2.8	3.0
Cholangiocarcinoma(C0206698)	Colonoscopy(C0009378)	1.0	1.3

which two variables correspond in their ranks [22]. It calculates how much a variable is statistically dependent on another, i.e., if one variable increases, how much the other variable tends to increase. The increase in the variable score may or may not be linear. If the score of one variable increases and the other decreases, the correlation coefficient would be negative. The correlation value ranges from a maximum of +1.00 through 0.00 to -1.00. The + sign indicates a positive correlation (the scores on one variable increase as the scores of the other variable increase). The - sign indicates a negative correlation (the scores on one variable increase, the scores on the other variable decrease). From the data sets X and Y, the N raw data scores (X_i, Y_i) are converted into ranks (x_i, y_i). The assignment of the rank is done by sorting the values in descending or ascending order. If two values are tied, then same rank which is equal to mean of their positions (if they were not same) is assigned to them. For example in Table 4.4, the same values (30), get the same rank.

Table 4.4: Spearman's Rank Assignment

X_i	Position in descending order	Rank
10	5	5
15	4	4
30	3	$2.5(\frac{3+2}{2})$
30	2	$2.5(\frac{3+2}{2})$
100	1	1

If the data has tied ranks, the correlation coefficient value is calculated by the formula :

$$\rho = \frac{\Sigma_i(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_i(x_i - \bar{x})^2 \times \Sigma_i(y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are mean values of the data sets X and Y respectively. Or else if the data does not have tied ranks following simplified formula can be used to calculate the correlation :

$$\rho = 1 - \frac{6*\Sigma(d_i)^2}{n*(n^2-1)}$$

where d_i is difference in paired ranks and n is total number of scores.

Let us calculate the correlation between marks obtained by a student in his painting class versus the marks obtained in his statistics class. Table 4.5 shows an example with two data sets which show the marks of five students.

Table 4.5: Spearman's Correlation Coefficient Example

Student	X_i (Marks in Painting class)	Y_i (Marks in Statistics class)
David	19	15
Brian	20	11
Raj	13	10
Lucy	14	14
Kate	15	20

Now, let us assign ranks to these positions. Table 4.6 shows the data sets with rank assigned to each variable value. It also shows the values of d_i and d_i^2.

Table 4.6: Spearman's Correlation Coefficient Calculation

Student	X_i	Y_i	Rank x_i	Rank y_i	d_i	(d_i^2)
David	19	15	2	2	0	0
Brian	20	11	1	4	-3	9
Raj	13	10	5	5	0	0
Lucy	14	14	4	3	1	1
Kate	15	20	3	1	2	4

Thus substituting $\Sigma(d_i)^2 = 14$ and $n = 5$ in the simplified formula for Spearman's correlation coefficient, the final value for coefficient is 0.3, which shows that the correlation between the marks obtained by student in painting and in statistics is low.

As Spearman's correlation coefficient calculation is based on the ranks instead of actual values of the variables, it is suitable for comparing sets of relatedness values with different ranges. The range of the relatedness score (lowest - highest) generated by the HSO algorithm is different than the range of the scores given by human judgments. For

example, the range of the relatedness values generated by program is (0 - 20), whereas the range of the MiniMayoSRS.physicians and MiniMayoSRS.coders gold standard is (0 - 10). To conclude, Spearman's correlation coefficient is a useful and suitable way to determine the ranked correlation between semantic relatedness values assigned by HSO algorithm and by human judgments, as it gives us a numerical measure of the amount of association between two sets of scores.

Chapter 5

Experimental Results

The application of the HSO on UMLS data led to various interesting results and observations. We present a set of experiments performed on the data sets (MiniMayoSRS, MayoSRS, UMNSRS_reduced_rel and UMNSRS_reduced_sim) explained in the previous chapter. We put forth a list of hypotheses, and then try to validate them by performing set of experiments necessary and finally present the conclusions.

The validity of the hypothesis is evaluated by calculating Spearman's correlation values for experimental data sets by comparing against the gold standards. Semantic relatedness values on a scale ranging from 0 to 20 are assigned by the algorithm to the concept pairs from the experimental data sets. If there is no path found between concepts, a semantic relatedness of -1 is assigned to the concept pairs. We also compare the results to the path measure provided by the UMLS::Similarity package, which calculates the semantic relatedness between concepts using the path distance between them.

As discussed in the Section 2.3, the UMLS consists of various source vocabularies, SNOMEDCT being the largest of them. As the experimental data sets were developed with concepts from SNOMEDCT, it is used as a primary source vocabulary for most of the experiments. As MSH is another popular vocabulary, it is used as comparison for the results obtained with SNOMEDCT. Experiments are performed for MiniMayoSRS test set on MSH vocabulary as MiniMayoSRS.msh.gold standard is available for evaluation purposes. Results obtained for MSH vocabulary are tabulated along with SNOMEDCT for most of the experiments, to observe HSO's application on different UMLS sources' graph structures.

All the results presented in this chapter, are calculated using the latest release of UMLS (at the time of experimentation), 2012AA. The cost of each upward and downward link is assumed to be equal to 1 and the cost of one horizontal link is assumed to be 2. This difference in cost makes sure that the horizontal paths are more expensive than up and down paths, as per HSO's suggestion. As the original HSO algorithm does not define a standard cost difference between up/down and horizontal links, we initially assume that the cost of traveling along one horizontal link is twice the cost of traveling up or down one link. Along with the semantic relatedness values and the correlation values obtained for different configurations with HSO, values of N are also specified whenever needed. The value of N is defined as the number of CUI pairs who have non-negative semantic relatedness values and are used by Spearman's correlation coefficient calculation.

Now that we have discussed the default configuration details and general scoring mechanism for semantic relatedness calculation, we present the list of hypotheses for which we designed the experiments.

1. The HSO measure when applied with only up (PAR relation) and down (CHD relation) vectors is equivalent to the shortest path measure implemented by the UMLS::Similarity package.

2. All relations and attributes from SNOMEDCT vocabulary except PAR and CHD can be used to represent horizontal links.

3. The Addition of horizontal relations and attributes selected by hypothesis 2 improves the correlation to the gold standards.

4. When the cost of traveling one Horizontal link is greater than the cost of one vertical link, the correlation with the gold standards is improved.

5. All possible allowable path patterns described by the HSO measure can be observed in the SNOMEDCT vocabulary, as it is a sufficiently large vocabulary.

6. If the path vectors in an allowable path are restricted in length, it correlates more with gold standard values, by reducing the number of false positives.

7. Allowing two direction changes in an allowable path between medical concepts aids in improving the correlation with gold standards.

5.1 Hypothesis 1: The HSO measure when applied with only up (PAR relation) and down (CHD relation) vectors is equivalent to the shortest path measure implemented by the UMLS::Similarity package.

Experiment 1: We evaluated the HSO algorithm by using PAR relation as upward link and CHD relation as downward link, using both the SNOMEDCT and MSH vocabulary.[1] We then compared Spearman's correlation values to the correlation values obtained by using the path measure implemented by UMLS::Similarity package. We used the default allowable path patterns of HSO, using the regular expression shown in Figure 3.5 to perform the experiments.

Observations and analysis : Tables A.4, A.5, A.6 and A.7 show the semantic relatedness (SR) values obtained for the MiniMayoSRS and MayoSRS data sets respectively, using the default configuration shown in Figure 5.1. The default configuration finds allowable paths that consist of up and down vector patterns where, PAR relation is used as a U link. As the graph of concepts is bi-directional, considering U links also implicitly includes D links using CHD relation. For example, if *Plant structure* is PAR of *Flowers*, then conversely, *Flowers* is CHD of *Plant structure*. The cost of each upward and downward link is equal to 1.

The baseline Spearman's correlation values obtained using the default configuration for both SNOMEDCT and MSH are shown in Table 5.1.

```
SAB :: include SNOMEDCT
REL :: include PAR
DIR :: include U
```

Figure 5.1: Default SNOMEDCT configuration file

[1] Experiment is performed using Perl package WebService::UMLSKS::Similarity.

```
SAB :: include MSH
REL :: include PAR
DIR :: include U
```

Figure 5.2: Default MSH configuration file

Table 5.1: Baseline Spearman's Correlation for SNOMEDCT and MSH (REL:PAR, DIR:U)

Data set(Size)	Correlation (SNOMEDCT)	Correlation (MSH)
MiniMayoSRS.physicians(29)	0.3133 (N = 26)	0.5143 (N = 22)
MiniMayoSRS.coders(29)	0.5102 (N = 26)	0.5403 (N = 22)
MayoSRS.gold(101)	0.1743 (N = 87)	NA
UMNSRS_reduced_rel.gold(430)	0.2945 (N = 363)	NA
UMNSRS_reduced_sim.gold(401)	0.5247 (N = 342)	NA

Comparison with path measure of UMLS-Similarity Table 5.2 shows the comparison between Spearman's correlation values obtained by HSO and the path measure of UMLS::Similarity, a CPAN module that calculates the semantic similarity between concepts using SNOMEDCT as a source vocabulary. The HSO algorithm uses PAR relation as upward vector and CHD relation as downward vector. Table 5.3 shows the comparison between the Spearman's correlation values obtained HSO and path measure of UMLS-Similarity using MSH as a source vocabulary. The correlation values calculated by UMLS::Similarity path measure are seen to be comparable to the HSO measure using up and down links.

Table 5.2: Comparison between correlation values of HSO default configuration (SAB: SNOMEDCT) and UMLS-Similarity path measure

Data set	HSO (U/D)	UMLS-Similarity path
MiniMayo.physicians(29)	0.3133 (N = 26)	0.3430 (N = 26)
MiniMayo.coders(29)	0.5102 (N = 26)	0.4741 (N = 26)
Mayo.gold(101)	0.1743 (N = 86)	0.1612 (N = 87)
UMNSRS.rel.gold(430)	0.2945 (N = 363)	0.2918 (N = 366)
UMNSRS.sim.gold(401)	0.5247 (N = 342)	0.5188 (N = 344)

Table 5.3: Comparison between correlation values of HSO default configuration (SAB:MSH) and UMLS-Similarity path measure

Data set	HSO (U/D)	UMLS-Similarity path
MiniMayo.msh.physicians(25)	0.5143 (N = 22)	0.3754 (N = 25)
MiniMayo.msh.coders(25)	0.5403 (N = 22)	0.4277 (N = 25)

Thus, the above comparison shows that when the HSO measure is used with only 'PAR' and 'CHD' relations i.e., upward and downward links, it is functionally equivalent to the path measure in UMLS::Similarity. This provides us with baseline correlation values that function as a point of comparison to measure the effect of further experiments such as adding horizontal links to path, experimenting with the allowed number of direction changes and restricting the allowed movement in each direction of the path, etc.

Conclusion: The experimental results support the hypothesis.

5.2 Hypothesis 2: All relations attributes from SNOMEDCT vocabulary from relation RO (other relations) can be used to represent horizontal links.

The SNOMEDCT widely uses defining relations which are used for defining a concept using its relationships with neighboring concepts [12]. The defining characteristics consist of 'ISA' relations (PAR/CHD relations) and 'Defining attribute relationships' (Other relations such as RO, RB, RN, etc). The 'ISA' relation attribute which is linked to 'PAR' relation is used for up and down links in this thesis work as described in previous section. Thus, the 'Defining attribute relationships' specifically from the relation 'RO' emerge as a useful addition to the set that represents the horizontal links between the concepts. 'RO' relation consists of largest variety and number of relation attributes.

As described in Background chapter, some defining attribute relationships connect all the source concepts to same destination concept using qualifier relations such as 'severity', 'episodicity', 'priority', 'clinical course', etc or temporal relations such as 'may_be_a', 'moved_from', 'replaced_by', 'was_a', etc. These qualifier or temporal relations are beneficial in knowing characteristics of concepts and may be used by UMLS to keep meta-data information about them. For example, a concept 'common cold' is connected to 'Severities' by relation attribute 'severity_of' to know how severe 'common cold' is. Furthermore, the relation attribute 'Severity_of' from relation RO, connects all the source concepts such as 'common cold', 'pneumonia', etc to only one destination concept called 'Severities'. Similarly attributes such as 'episodicity_of', 'priority_of','has_clinical_course', etc are other qualifying relations.

Experiment 2: We configured the HSO algorithm using PAR relation as upward link, CHD relation as downward link and all attributes (including temporal and qualifier attributes) from other relations (RO) as horizontal links, for SNOMEDCT vocabulary.[2]
We then studied the semantic relatedness values and path lengths obtained for the concept pairs from MiniMayoSRS data set. We used default allowable path patterns suggested by HSO using a regular expression shown in Figure 3.5 to perform the experiments.

[2] Experiment is performed using Java package WebService::UMLSKS::Similarity::Java.

Observations and analysis : Table 5.4 shows the effect on semantic relatedness values for CUI pairs from MiniMayoSRS data set, when all the relation attributes (including temporal and qualifying attributes) from relation 'RO' are included in the allowed set of H links.

Table 5.4: Set of CUIs from MiniMayoSRS (SAB : SNOMEDCT, REL : PAR,RO, DIR : U,H)

Source CUI	Destination CUI	SR	Qualifier or Temporal RELA
Renal failure (C0035078)	Kidney failure (C0035078)	20.00	No
Abortion (C0156543)	Miscarriage (C0000786)	19.00	No
Congestive heart failure (C0018802)	Pulmonary edema (C0034063)	16.00	Yes
Diarrhea (C0011991)	Stomach cramps (C0344375)	16.00	Yes
Mitral stenosis (C0026269)	Atrial brillation (C0004238)	16.00	Yes
Pulmonary embolus (C0034065)	Myocardial infarction (C0027051)	16.00	Yes
Carpal tunnel syndrome (C0007286)	Osteoarthritis (C0029408)	16.00	Yes
Rheumatoid arthritis (C0003873)	Lupus (C0409974)	16.00	Yes
Peptic ulcer disease (C0030920)	Myopia (C0027092)	16.00	Yes
Lymphoid hyperplasia (C0333997)	Laryngeal cancer (C0007107)	16.00	Yes
Depression (C0011581)	Cellulitis (C0007642)	16.00	Yes
Multiple sclerosis (C0026769)	Psychosis (C0033975)	16.00	Yes
Xerostomia (C0043352)	Alcoholic cirrhosis (C0023891)	16.00	Yes
Diabetes mellitus (C0011849)	Hypertension (C0020538)	16.00	Yes
Appendicitis (C0003615)	Osteoporosis (C0029456)	16.00	Yes
Cortisone (C0010137)	Total knee replacement (C0086511)	14.00	Yes
Pulmonary brosis (C0034069)	Lung cancer (C0242379)	12.75	No
Brain tumor (C0006118)	Intracranial hemorrhage (C0151699)	12.75	No
Stroke (C0038454)	Infarct (C0021308)	12.00	No
Antibiotic (C0003232)	Allergy (C0020517)	11.25	No
Acne (C0702166)	Syringe (C0039142)	7.5	No

As it can be observed that 16 out of 21 CUI pairs get connected to each other through a qualifier or temporal relation attributes. Out of 21 concept pairs 14 concept pairs are assigned same relatedness value of 16.00, as they all are connected through path length of 4. Shortest allowable paths from each of these 14 concept pairs, have following allowable path pattern :

Source concept (H) - Common Associated concept (H) - Destination Concept

where, H link corresponds to a qualifier or temporal relation attribute such as 'episodicity_of' or 'may_be_a' and has cost of 2. The qualifier and temporal relationships

do not relate two concepts based on their meaning. So, when these relation attributes are included in allowed set of H links, any two concepts that are connected to the concepts such as 'Severities' and 'Episodicities', are connected to each other, even if they are semantically far away from each other. Any two entities whose episodicity can be measured are connected through common neighbor 'Episodicities' and have incorrect semantic relatedness value. This shows that even though most of defining attribute relations are useful in the calculation of semantic relatedness, not all relations contribute to the correct calculation of relatedness value between a concept pair. Thus identifying if the relationship is a qualifier or temporal helps us refine the set of allowable relation for H links. Another information used for refinement of relation set was the frequency of the relation and its attributes. If the relation or attribute appears quite more times as compared to other relations and attributes in the source vocabulary, it will allow HSO to explore large number of related neighbors and find more interesting paths.

Thus a set of relations attributes is formed using following criteria :

- Frequency of the relation attribute
- Is the relation attribute useful, i.e., it is not a temporal or qualifier relation attribute.

Table 5.5 shows top attributes from 'RO' relation along with their frequencies and whether the attribute is useful for the calculation of semantic relatedness or not. The relation attributes added to relate the concepts based on their actual meaning are used to form the allowed set of relation attributes for 'RO' relation.

Choosing top frequency useful relation attributes from Table 5.5, set of 26 relation attributes for RO relation is defined as -
{finding_site_of, has_finding_site, mapped_to, mapped_from, method_of, has_method, associated_morphology_of, has_associated_morphology, has_direct_procedure_site, direct_procedure_site_of, active_ingredient_of, has_active_ingredient, has_causative_agent, causative_agent_of, access_of, has_access, has_component, component_of, has_dose_form, dose_form_of, has_definitional_manifestation, definitional_manifestation_of, uses_device, device_used_by, interprets, is_interpreted_by}

Table 5.5: Top Attributes' from RO relation

Attribute	Useful Horizontal relation?	Frequency
episodicity_of	No	82243
has_episodicity	No	82243
has_clinical_course	No	81791
clinical_course_of	No	81791
severity_of	No	81737
has_severity	No	81737
finding_site_of	Yes	69702
has_finding_site	Yes	69702
method_of	Yes	56233
has_method	Yes	56233
has_priority	No	51337
priority_of	No	51337
associated_morphology_of	Yes	50003
has_associated_morphology	Yes	50003
has_direct_procedure_site	Yes	29948
direct_procedure_site_of	Yes	29948
inverse_may_be_a	No	29610
may_be_a	No	29610
access_of	Yes	28806
has_access	Yes	28806
is_interpreted_by	Yes	23794
interprets	Yes	23794
has_active_ingredient	Yes	18676
active_ingredient_of	Yes	18676
has_causative_agent	Yes	16924
causative_agent_of	Yes	16924
has_laterality	No	16194
laterality_of	No	16194
moved_to	No	14451
moved_from	No	14451
has_dose_form	Yes	10971
dose_form_of	Yes	10971
has_component	Yes	8738
component_of	Yes	8738
has_indirect_procedure_site	Yes	7812
indirect_procedure_site_of	Yes	7812
occures_in	No	7561
has_occurance_in	No	7561

Conclusion: Thus, experimental results show that selection of horizontal relation attributes is necessary, which disagrees with the hypothesis. The relation set that can be used as horizontal relations is shown in Figure 5.3.

{finding site of, has finding site, mapped to, mapped from, method of, has method, associated morphology of, has associated morphology, has direct procedure site, direct procedure site of, active ingredient of, has active ingredient, has causative agent, causative agent of, access of, has access, has component, component of, has dose form, dose form of, has definitional manifestation,definitional manifestation of, uses device, device used by, interprets, is interpreted by}

Figure 5.3: Selected H relations and attributes

5.3 Hypothesis 3: The addition of horizontal relations and attributes selected by hypothesis 2 improves the correlation to the gold standards.

Along with using parent (PAR) and child (CHD) relations as up and down vectors, all the relation attributes suggested by hypothesis 2, are used as horizontal links for the calculation of semantic relatedness. Hypothesis 2 suggests attributes from RO (other relations) relation, which is the most widely used horizontal relation in SNOMEDCT. Other useful horizontal relations are RB (Broader relation) and RN (Narrower relation). The relations 'RB' and 'RN' represent the broader and narrower relation types between concepts and have 'has_part' and 'part_of' relation attributes as most frequently occurring attributes. When HSO measure was introduced by Hirst and St.Onge for WordNet, 'PART_OF' relationships were used to represent horizontal links. Thus, inspired by HSO's implementation for WordNet, relations 'RB' and 'RN' with attributes 'has_part' and 'part_of' are also considered for representing H links, along with selected attributes from RO relation from hypothesis 2. Finally, Figure 5.4 shows the resultant configuration file that uses RO, RB and RN relations with 26 relation attributes to represent horizontal paths, while performing experiments with horizontal relations.

```
SAB :: include SNOMEDCT
REL :: include PAR,RO,RB,RN
DIR :: include U,H,H,H
RELA :: include RO-finding_site_of,RO- has_finding_site,RB-has_part,
RN-part_of,RO-mapped_to,RO-mapped_from,RO-method_of,RO-has_method,RO-
associated_morphology_of,RO-has_associated_morphology,RO-has_direct_procedure_site,RO-
direct_procedure_site_of,RO-active_ingredient_of,RO-has_active_ingredient,RO-
has_causative_agent,RO-causative_agent_of,RO-access_of,RO-has_access,RO-has_component,
RO-component_of,RO-has_dose_form, RO-dose_form_of,RO-has_definitional_manifestation, RO-
definitional_manifestation_of,RO-uses_device,RO-device_used_by,RO-interprets,RO-
is_interpreted_by
```

Figure 5.4: Configuration file with selected H relations and attributes

To efficiently perform the experiments with large number of relations and relation attributes shown in Figure 5.4 as horizontal links, big subsets are used as experimental

data sets. Table 4.3 and Table A.8 with Table A.9 tabulate the big subsets and their keys for MiniMayoSRS and MayoSRS test sets respectively. Experiment 3 and 4 present the results obtained by experimenting with these big subsets. As the big subsets do not contain all the concept pairs from original data sets, we re-calculated the baseline correlation for big subsets by performing experiment with default configuration.

Experiment 3: We configure the HSO algorithm using default configuration where PAR relation is used as upward link and CHD relation is used as downward link for SNOMEDCT vocabulary.[3] We assign the semantic relatedness values for concept pairs from the big subsets and then calculate the baseline correlation values for them. We used default allowable path patterns suggested by HSO, using a regular expression shown in Figure 3.5 to perform the experiments.

Observations and analysis : Table A.10 and A.11 with Table A.12 show the semantic relatedness (SR) values obtained for the subsets of MiniMayoSRS and MayoSRS data set respectively using the concept of allowable paths for the up and down vector patterns, where PAR relation is used as a U link.

Similar to the experiments with original data sets, semantic relatedness values are obtained using the default configuration file as shown in Figure 5.1. As the graph of concepts is bi-directional, considering U links also implicitly includes D links. Table 5.6 shows the Spearman's correlation values for all big subsets using the default configuration. These correlation values serve as a baseline for further experiments performed on big subsets.

Experiment 4: We configured the HSO algorithm using PAR relation as upward link, CHD relation as downward link, RB(Broader relation), RN(Narrower relation) and chosen relation attributes set from RO relation as horizontal links, for SNOMEDCT vocabulary.[4] The cost of each horizontal link is equal to two. We then studied the semantic relatedness values and path lengths obtained for the concept pairs from big subsets. We also compare Spearman's correlation values of this experiment with baseline correlation values from Experiment 3. We used default allowable path patterns

[3] Experiment is performed using Java package WebService::UMLSKS::Similarity::Java.
[4] Experiment is performed using Java package WebService::UMLSKS::Similarity::Java.

Table 5.6: Correlation values using default configuration (SAB : SNOMEDCT)

Data set	(U/D)
MiniMayo.subset.coders (23)	0.6262 (N = 23)
MiniMayo.subset.physicians (23)	0.3462 (N = 23)
Mayo.subset.gold (84)	0.2553 (N = 82)
UMNSRS.rel.subset.gold (366)	0.3082 (N = 346)
UMNSRS.sim.subset.gold (341)	0.5214 (N = 323)

suggested by HSO, using a regular expression shown in Figure 3.5 to perform the experiments.

Observations and analysis : The semantic relatedness values obtained for the MiniMayoSRS big subset are tabulated in Table A.16 and relatedness values for MayoSRS big subset are tabulated in Table A.17 and A.18.

We observed interesting effects on the semantic relatedness values, path lengths and path costs between the concept pairs, after adding horizontal relations and attributes. Table 5.7 shows the number of CUI pairs from the big subsets whose semantic relatedness value increased after the addition of horizontal links. As it can be observed, the path lengths and costs of around 10% of the CUI pairs on an average, decreased than the path lengths and costs obtained by using only up and down links.

Table 5.7: Number of CUI pairs for which SR value increased after adding H links using cost of H - 2 (SAB : SNOMEDCT)

Data set	Number of CUI pairs
MiniMayo.subset.physicians (23)	3
MiniMayo.subset.coders (23)	3
Mayo.subset.gold (84)	15
UMNSRS.rel.subset.gold (366)	23
UMNSRS.sim.subset.gold (341)	29

Figure 5.5 shows two examples of how the path between concept pairs changed after adding the horizontal relations and attributes. The first example shows the path between C0003873 (Rheumatoid Arthritis) and C0003904 (Arthroscopy). The path obtained using only up and down relations as shown in first column, travels up to the root of the vocabulary and connects the CUI pair with a path with cost of 11. Whereas, the path after adding H relations connects the CUI pair with more meaningful path with cost equal to 4. The second example also shows similar reduction in path cost after adding H relations between C0006121 (Brain Stem) and C1269897 (Entire cranial nerve). These examples also show the use of selected relations attributes 'has_finding_site' and 'direct_procedure_site_of' from RO relation, as H links.

Table 5.8 shows the comparison between the correlation values obtained for big subsets with and without horizontal relations, when the cost of each H link is two. As it can be observed in the correlation comparisons, the correlation values improve after the addition of horizontal relations.

Table 5.8: Comparison between correlation values of default configuration and configuration with H relations, cost of H - 2 (SAB : SNOMEDCT)

Data set	(U/D)	(U/D & H)
MiniMayo.subset.coders (23)	0.6262	0.8045
	(N = 23)	(N = 23)
MiniMayo.subset.physicians (23)	0.3462	0.5085
	(N = 23)	(N = 23)
Mayo.subset.gold (84)	0.2553	0.3827
	(N = 82)	(N = 83)
UMNSRS.rel.subset.gold (366)	0.3082	0.3053
	(N = 346)	(N = 346)
UMNSRS.sim.subset.gold (341)	0.5214	0.5128
	(N = 323)	(N = 323)

Source and Destination	Path Before adding H relations	Path after adding H relations
C0003873 (Rheumatoid Arthritis) and C0003904 (Arthroscopy)	C0003873(Rheumatoid Arthritis)(U)[PAR - inverse_isa] -> C0003864(Arthritis)(U)[PAR - inverse_isa] -> C1285331(Inflammation of specific body organs)(U)[PAR - inverse_isa] -> C1285332(Inflammation of specific body structures or tissue)(U)[PAR - inverse_isa] -> C1290853(Disorder by body site)(U)[PAR - inverse_isa] -> C0012634(Disease)(U)[PAR - inverse_isa] -> C0037088(Signs and Symptoms)(U)[PAR - inverse_isa] -> C2720507(SNOMED CT Concept (SNOMED RT+CTV3))(D)[CHD - isa] -> C0184661(Interventional procedure)(D)[CHD - isa] -> C1285536(Procedure categorized by device involved)(D)[CHD - isa] -> C0014245(Endoscopy (procedure))(D)[CHD - isa] -> C0003904(Arthroscopy) Semantic relatedness : 6.75 Path cost : 11 Number of changes in direction : 1	C0003873(Rheumatoid Arthritis)(H)[RO - has_finding_site] -> C0022417(Joints)(H)[RO - direct_procedure_site_of] -> C0003904(Arthroscopy) Semantic relatedness : 16.00 Path cost : 4 Number of changes in direction : 0
C0006121 (Brain Stem) and C1269897 (Entire cranial nerve)	C0006121(Brain Stem)(U)[PAR - inverse_isa] -> C0459385(Brain tissue)(U)[PAR - inverse_isa] -> C0445620(Brain part)(U)[PAR - inverse_isa] -> C0504215(Organ part)(U)[PAR - inverse_isa] -> C0229983(Body organ structure)(D)[CHD - isa] -> C1280836(Entire body organ)(D)[CHD - isa] -> C1280541(Entire nerve)(D)[CHD - isa] -> C1269897(Entire cranial nerve) Semantic relatedness : 9.75 Path cost : 7 Number of changes in direction : 1	C0006121(Brain Stem)(H)[RO - has_finding_site] -> C0393799(Miller Fisher Syndrome)(H)[RO - has_finding_site] -> C0010268(Cranial Nerves)(D)[CHD - isa] -> C1269897(Entire cranial nerve) Semantic relatedness : 11.25 Path cost : 5 Number of changes in direction : 1

Figure 5.5: Examples with updated path after adding H relations

Experiment 5: We configured the HSO algorithm using PAR relation as upward link, CHD relation as downward link, SIB (Sibling relation) as horizontal links, for MSH vocabulary.[5] The cost of each horizontal link is two. We then studied the semantic relatedness values and path lengths obtained for the concept pairs from MiniMayoSRS data set. We also compare the Spearman's correlation values of this experiment with baseline correlation values from Experiment 1. We used default allowable path patterns suggested by HSO, using a regular expression shown in Figure 3.5 to perform the experiments.

Observations and analysis : The 'SIB' (sibling) relation is chosen as shown in Figure 5.6 for representing H links as it is most frequent relation that occurs in MSH. Results show that selection of SIB as H relation for MSH vocabulary leads to interesting paths between the concept pairs from data sets. Table 5.9 presents Spearman's correlation coefficient values for MiniMayoSRS test set for MSH. The correlation values obtained after adding SIB (sibling relation) as H link, can be compared against the baseline correlation values obtained by using only U and D links. Similar results for SNOMEDCT vocabulary, addition of H relations improves the correlation with gold standards even in case of MSH vocabulary.

```
SAB :: include MSH
REL :: include PAR,SIB
DIR :: include U,H
```

Figure 5.6: MSH configuration file with SIB relation as H link

Table 5.9: Spearman's Correlation Values (SAB:MSH, REL:PAR,SIB DIR:U,H)

Data set	With H relations	Baseline
MiniMayoSRS.physicians(25)	0.6575	0.5143
	(N = 22)	(N = 22)
MiniMayoSRS.coders(25)	0.6249	0.5403
	(N = 22)	(N = 22)

[5] Experiment is performed using Java package WebService::UMLSKS::Similarity::Java.

We now summarize the observations made after adding horizontal relations and attributes to the calculation of semantic relatedness:

- A shorter path is found using H links as compared to path found using U and D links in case of CUI pairs for which H links were part of final path. Thus the semantic relatedness value increases in case of such CUI pairs as the result of shorter path.
- Paths which connect the concepts through root concept of the source vocabulary (when only U/D links are used) are substituted by more meaningful path after the addition of H relations.
- Addition of H relations allows the algorithm to assign semantic relatedness values that correlate with gold standards.
- Along with SNOMEDCT vocabulary adding H relations also improved the Spearman's correlation values for MSH vocabulary.

Conclusion: Experimental results support the hypothesis, as addition of horizontal relations and attributes selected by hypothesis 2 improves the correlation to the gold standards.

5.4 Hypothesis 4: When the cost of traveling one Horizontal link is greater than the cost of one vertical link, the correlation to the gold standards is improved.

As HSO suggests, the more you divert from the original concept horizontally, the more you go away from it's meaning. In both the upward and downward links, the concepts do not digress a lot from the meaning of original concept. Therefore, you remain close to the context of the original concept by traveling either towards more general concept or more specific concept. But, when horizontal links which correspond to relations like aggregation, associations, etc., are followed, the concepts tend to digress from the meaning of original concept. Thus, H links are considered to be expensive as compared to U and D links in original HSO algorithm. By performing following experiments we try to find if H links should be penalized and if yes, what is the correct degree of penalty.

Experiment 6: We configured the HSO algorithm using PAR relation as upward link, CHD relation as downward link, RB(Broader relation), RN(Narrower relation) and chosen relation attributes set from RO relation as horizontal links, for SNOMEDCT vocabulary.[6] We perform experiments with cost of H link = 1, followed by experiments with cost of H link = 3. Results with cost of H = 1, 2 and 3 are compared against the baseline for big subsets and with each other, to study the changes in values of semantic relatedness, path costs and Spearman's correlation coefficient. We used default allowable path patterns suggested by HSO, using a regular expression shown in Figure 3.5 to perform the experiments.

Observations and analysis : To observe the effect of equalizing the cost of up and down link to cost of horizontal link, we calculated the semantic relatedness values for CUI pairs from big subsets, by setting the cost of up, down and horizontal link to one. The semantic relatedness values obtained for the MiniMayoSRS big subset are tabulated in Table A.19 and relatedness values for MayoSRS big subset are tabulated in Table A.20 and A.21.

Table 5.10 shows the comparison between the correlation values obtained for big

[6] Experiment is performed using Java package WebService::UMLSKS::Similarity::Java.

subsets with and without horizontal relations, when the cost of each H link is 1. As it can be observed that the correlation values drop significantly when the cost of traveling each H link is equal to cost of traveling a U or D link. This agrees with HSO's suggestion that the cost of H link should be greater than a U or D link. This shows that even in case of UMLS graph, the more you travel from the original concept horizontally, the more you go away from it's meaning. Thus, similar to WordNet, a English vocabulary graph, HSO's suggestion stands true for UMLS, a medical vocabulary graph. Thus, H links should be considered to be expensive as compared to U and D links.

Table 5.10: Comparison between correlation values of default configuration and Configuration with H relations, cost of H - 1 (SAB : SNOMEDCT)

Data set	(U/D)	(U/D & H)
MiniMayo.subset.coders (23)	0.6262	0.6184
	(N = 23)	(N = 23)
MiniMayo.subset.physicians (23)	0.3462	0.4041
	(N = 23)	(N = 23)
Mayo.subset.gold (84)	0.2553	0.2105
	(N = 82)	(N = 83)
UMNSRS.rel.subset.gold (366)	0.3082	0.2109
	(N = 346)	(N = 346)
UMNSRS.sim.subset.gold (341)	0.5214	0.4017
	(N = 323)	(N = 323)

As we have now confirmed that the cost of H link should be greater than that of U/D link, we study the effect of degree of penalty of H links. We find the results for big subsets when cost of one H link is thrice the cost of one U/D link. The semantic relatedness values obtained when the cost of H link is equal to 3, for the MiniMayoSRS big subset are tabulated in Table A.13 and relatedness values for MayoSRS big subset are tabulated in Table A.14 and Table A.15. Table 5.12 shows the comparison between the correlation values obtained against the baseline correlations for the big subsets.

These experimental results show us the effect of heavily penalizing horizontal links than upward or downward link as each horizontal link's cost is three times the cost of a up or down link. It is observed during these experiments that as the algorithm tries to find shortest allowable path between the source and destination concept, it prefers a

path with up and down relations over the path with H links in most of the cases. When path with H links is found between a concept pairs, it is replaced by algorithm by a shorter allowable path using up and down links. Thus the use of H links in finding the shortest allowable paths is suppressed due to high cost. Figure 5.7 shows an example of how the algorithm finds different paths during the path search and finally settles down to the shortest allowable path formed using up and down links only.

C0002962(Angina Pectoris) and Destination : C0070166(clopidogrel)

Source : C0002962(Angina Pectoris) and Destination : C0070166(clopidogrel)

Replacing path.
Old path: ShortestPath: noc=1, cost=16, pathDirection=HHHHHD,
path=[C0002962(Angina Pectoris), C0817096(Chest), C0198382(Repair of thoracogastric fistula), C0038351(Stomach), C1828441(Gastric ulcer induced by anti-platelet agent), C0085826(Antiplatelet Agents), C0070166(clopidogrel)],

new path: ShortestPath: noc=1, cost=8, pathDirection=UUUUDDDD,
path=[C0002962(Angina Pectoris), C1300028(Disorder characterized by pain), C0012634(Disease), C0037088(Signs and Symptoms), C2720507(SNOMED CT Concept (SNOMED RT+CTV3)), C0013227(Pharmaceutical Preparations), C0007220(Cardiovascular Agents), C0085826(Antiplatelet Agents), C0070166(clopidogrel)]

Final Path: C0002962(Angina Pectoris)(U)[PAR - inverse_isa] -> C1300028(Disorder characterized by pain)(U)[PAR - inverse_isa] -> C0012634(Disease)(U)[PAR - inverse_isa] -> C0037088(Signs and Symptoms)(U)[PAR - inverse_isa] -> C2720507(SNOMED CT Concept (SNOMED RT+CTV3))(D)[CHD - isa] -> C0013227(Pharmaceutical Preparations)(D)[CHD - isa] -> C0007220(Cardiovascular Agents)(D)[CHD - isa] -> C0085826(Antiplatelet Agents)(D)[CHD - isa] -> C0070166(clopidogrel)

Semantic Relatedness : 9.0
Final cost : 8
Final Changes in directions : 1

Figure 5.7: Path between C0002962(Angina Pectoris) and C0070166(clopidogrel)

As seen in this example, due to heavy penalty applied for a H link algorithm prefers a shorter path between C0002962(Angina Pectoris) and C0070166(clopidogrel) with cost of 8. Table 5.11 shows the number of CUI pairs for which an allowable path with horizontal relations was replaced by a shorter path with up and down relations. Thus, less number of CUI pairs are connected using H links when the cost of H link is increased from 2 to 3. Further Table 5.13 shows the comparison between the number of CUIs for which the semantic relatedness values increased after adding horizontal relation with cost = 3 versus cost = 2. It can be observed that frequency of such CUIs is decreased if we increase the cost of H link from 2 to 3. Finally, as it can be observed from Table 5.12, by adding horizontal relations to HSO with cost of H link = 3, there is less improvement

in the correlation with gold standards, as compared to the improvement when cost of H link = 2.

Table 5.11: Number of CUI pairs for which path with H link was replaced by shorter path with U/D links, cost of H - 3 (SAB : SNOMEDCT)

Data set	Cost(H) - 2
MiniMayo.subset.physicians (23)	6
MiniMayo.subset.coders (23)	6
Mayo.subset.gold (84)	10
UMNSRS.rel.subset.gold (366)	90
UMNSRS.sim.subset.gold (341)	93

Table 5.12: Comparison between correlation values of default configuration and Configuration with H relations, cost of H - 3 (SAB : SNOMEDCT)

Data set	(U/D)	(U/D & H)
MiniMayo.subset.coders (23)	0.6262	0.8031
	(N = 23)	(N = 23)
MiniMayo.subset.physicians (23)	0.3462	0.4976
	(N = 23)	(N = 23)
Mayo.subset.gold (84)	0.2553	0.3244
	(N = 82)	(N = 83)
UMNSRS.rel.subset.gold (366)	0.3082	0.3165
	(N = 346)	(N = 346)
UMNSRS.sim.subset.gold (341)	0.5214	0.5236
	(N = 323)	(N = 323)

Finally, we compare the results for all three costs for H links, i.e., cost = 1, 2 and 3. Table 5.14 shows the comparison between the semantic relatedness values obtained using the cost of H link = 1, 2 and 3. The semantic relatedness values are compared against the correlation values obtained by using only U and D relations. Table 5.15 compares the number of CUI pairs for which the semantic relatedness value increased after the addition of H relations for all three costs.

Table 5.13: Comparison between number of CUI pairs for which SR value increased after adding H links using cost of H link is 3 and 2 (SAB : SNOMEDCT)

Data set	Cost(H) - 3	Cost(H) - 2
MiniMayo.subset.physicians (23)	2	3
MiniMayo.subset.coders (23)	2	3
Mayo.subset.gold (84)	7	15
UMNSRS.rel.subset.gold (366)	3	23
UMNSRS.sim.subset.gold (341)	4	29

Table 5.14: Comparison between semantic relatedness values after adding H links using cost of H link is 3, 2 and 1 (SAB : SNOMEDCT)

Data set	U/D links	Cost(H)-3	Cost(H)-2	Cost(H)-1
MiniMayo.subset.coders (23)	0.6262	0.8031	0.8045	0.6184
	(N = 23)	(N = 23)	(N = 23)	(N = 23)
MiniMayo.subset.physicians (23)	0.3462	0.4976	0.5085	0.4041
	(N = 23)	(N = 23)	(N = 23)	(N = 23)
Mayo.subset.gold (84)	0.2553	0.3244	0.3827	0.2105
	(N = 82)	(N = 83)	(N = 83)	(N = 83)
UMNSRS.rel.subset.gold (366)	0.3082	0.3165	0.3053	0.2109
	(N = 346)	(N = 346)	(N = 346)	(N = 346)
UMNSRS.sim.subset.gold (341)	0.5214	0.5236	0.5128	0.4017
	(N = 323)	(N = 323)	(N = 323)	(N = 323)

Table 5.15: Comparison between number of CUI pairs for which SR value increased after adding H links using cost of H link is 3, 2 and 1 (SAB : SNOMEDCT)

Data set	Cost(H) - 3	Cost(H) - 2	Cost(H) - 1
MiniMayo.subset.physicians (23)	2	3	9
MiniMayo.subset.coders (23)	2	3	9
Mayo.subset.gold (84)	7	15	31
UMNSRS.rel.subset.gold (366)	3	23	130
UMNSRS.sim.subset.gold (341)	4	29	124

Out of the three cost values (1, 2 and 3) for H links, it can be observed that by penalizing horizontal relations moderately, i.e. cost of $H = 2$, semantic relatedness values correlate more to gold standards and considerable number of meaningful horizontal relations are used to calculate the semantic relatedness.

Conclusion: Experimental results agree with the hypothesis and show that when H links are penalized moderately, it leads to better calculation of semantic relatedness.

5.5 Hypothesis 5: All possible allowable path patterns described by the HSO measure can be observed in the SNOMEDCT vocabulary, as it is a sufficiently large vocabulary.

Experiment 7: We configured the HSO algorithm using PAR relation as upward link, CHD relation as downward link, RB(Broader relation), RN(Narrower relation) and chosen relation attributes set from RO relation as horizontal links. The cost of H link is assumed to be twice the cost of U/D link. We then searched for different allowable path patterns between the concept pairs from experimental data sets, using SNOMEDCT vocabulary.[7] We used default allowable path patterns suggested by HSO, using a pattern regular expression shown in Figure 3.5 to perform the experiments.

Observations and analysis : As suggested by HSO, when a path follows one of the allowable path patterns, it accurately describes the relatedness between two concepts. These path patterns were defined considering WordNet, an English vocabulary graph. We present the allowable path patterns obtained between medical concepts from UMLS SNOMEDCT vocabulary graph in the form of example CUI pairs. Figure 5.8 shows two allowable path patterns 'U+' and 'D+'. The path pattern 'U+' represents an allowed path formed by one or more number of upward vectors, whereas path pattern 'D+' represents an allowed path formed by one or more number of upward vectors. The detailed path with relations and attributes used in shortest allowable path for pattern 1 and pattern 2 are as follows : Pattern 1 : C0001962(Ethanol)(U)[PAR - inverse_isa] - C1690586(Alcohol agent)(U)[PAR - inverse_isa] - C0001975(Alcohols)
Pattern 2 : C0026946(Mycoses)(D)[CHD - isa] - C0276697(Infection by Ascomycetes)(D)[CHD - isa] - C0005716(Blastomycosis)

Figure 5.9 shows two allowable path patterns 'H+' and 'H+D+'. The path pattern 'H+' represents an allowed path formed by one or more horizontal vectors, whereas path pattern 'H+D+' represents an allowed path formed by one or more horizontal vectors followed by one or more downward vectors. The detailed path with relations

[7] Experiment is performed using both Perl and Java packages.

and attributes used in shortest allowable path for pattern 3 and pattern 4 are as follows : Pattern 3 : C0032285(Pneumonia)(H)[RO - associated_morphology_of] - C0021368(Inflammation)(H)[RO - has_associated_morphology] - C1827213(Herpes zoster subepithelial infiltrates)(H) [RO - has_associated_morphology] - C0332448(Infiltration) Pattern 4 : C0009319(Colitis)(H)[RO - associated_morphology_of] - C0021368(Inflammation)(H)[RO - has_associated_morphology] - C0393484(Rasmussen Syndrome)(H)[RO - has_definitional_manifestation] - C0036572(Seizures)(D)[CHD - isa] - C0014544(Epilepsy).

Figure 5.10 shows two allowable path patterns 'D+H+' and 'U+H+'. The path pattern 'D+H+' represents an allowed path formed by one or more downward vactors followed by one or more horizontal vectors, whereas path pattern 'U+H+' represents an allowed path formed by one or more upward vectors followed by one or more horizontal vectors. The detailed path with relations and attributes used in shortest allowable path for pattern 5 and pattern 6 are as follows : Pattern 5 : C0086543(Cataract)(D)[CHD - isa] - C0009691(Congenital cataract)(D)[CHD - isa] - C0268361(Osteogenesis imperfecta, recessive perinatal lethal, with microcephaly AND cataracts)(H)[RO - has_finding_site] - C0006104(Brain)(H)[RO - finding_site_of] - C0014544(Epilepsy) Pattern 6 : C0032285(Pneumonia)(U)[PAR - inverse_isa] - C1285331(Inflammation of specific body organs)(U)[PAR - inverse_isa] - C1285332(Inflammation of specific body structures or tissue) (H)[RO - has_finding_site] - C2711400(Anatomical or acquired body structure)(H)[RO - finding_site_of] - C0036690(Septicemia)

Figure 5.11 shows an allowable path pattern 'U+D+', which represents an allowed path formed by one or more upward vactors followed by one or more downward vectors. The detailed path with relations and attributes used in shortest allowable path for pattern 7 is as follows: Pattern 7 : C0011175(Dehydration)(U)[PAR - inverse_isa] - C0267995(Fluid volume disorder)(U)[PAR - inverse_isa] - C0267994(Disorder of fluid AND/OR electrolyte)(U)[PAR - inverse_isa] - C0025517(Metabolic Diseases)(U)[PAR - inverse_isa] - C0012634(Disease)(D)[CHD - isa] - C0028709(Nutrition Disorders)(D)[CHD - isa] - C0162429(Malnutrition)(D)[CHD - isa] - C0038187(Starvation)

Figure 5.12 shows an allowable path pattern 'U+H+D+', which represents an allowed path formed by one or more upward vactors followed by one or more horizontal

vectors and then followed by one or more downward vectors. The detailed path with relations and attributes used in shortest allowable path for pattern 8 is as follows: Pattern 8 : C0019270(Hernia)(U)[PAR - inverse_isa] - C0333056(protrusion)(U)[PAR - inverse_isa] - C0333010(Mechanical abnormality)(U)[PAR - inverse_isa] - C0332447(Morphologically abnormal structure)(H)[RO - has_associated_morphology] - C0011609(Drug Eruptions)(H)[RO - has_causative_agent] - C0013227(Pharmaceutical Preparations)(D)[CHD - isa] - C0014432(Enzyme Inhibitors)(D)[CHD - isa] - C0598272(DOPA decarboxylase inhibitor)(D)[CHD - isa] - C0006982(Carbidopa).

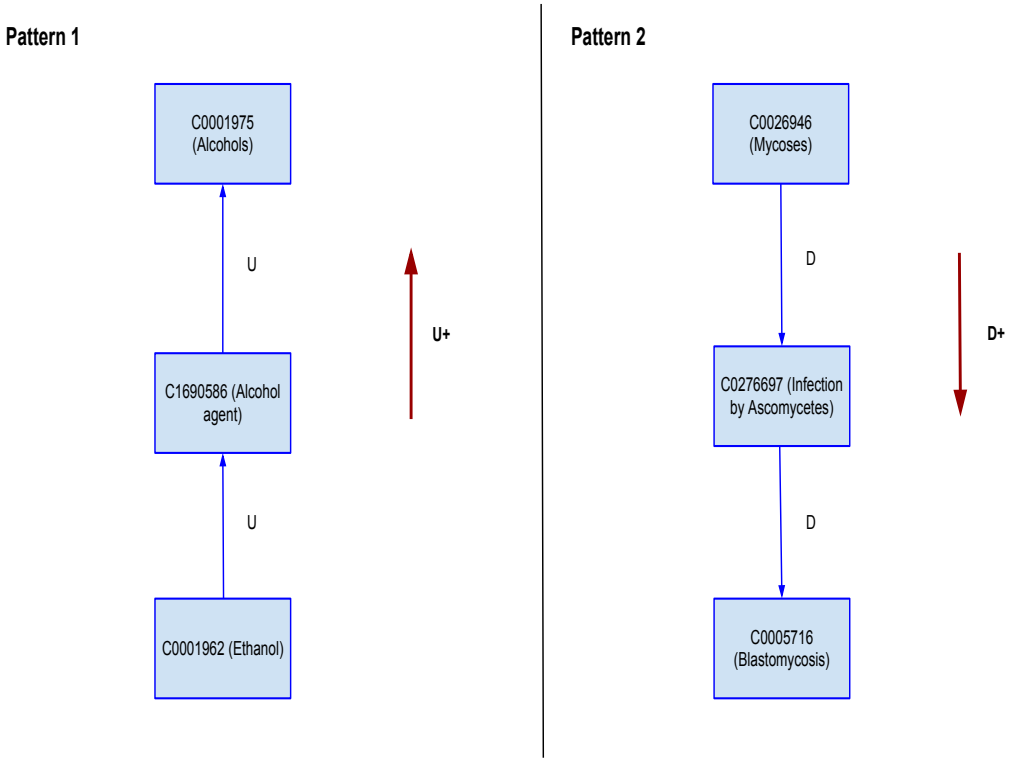

Figure 5.8: Allowable path pattern 1 and 2

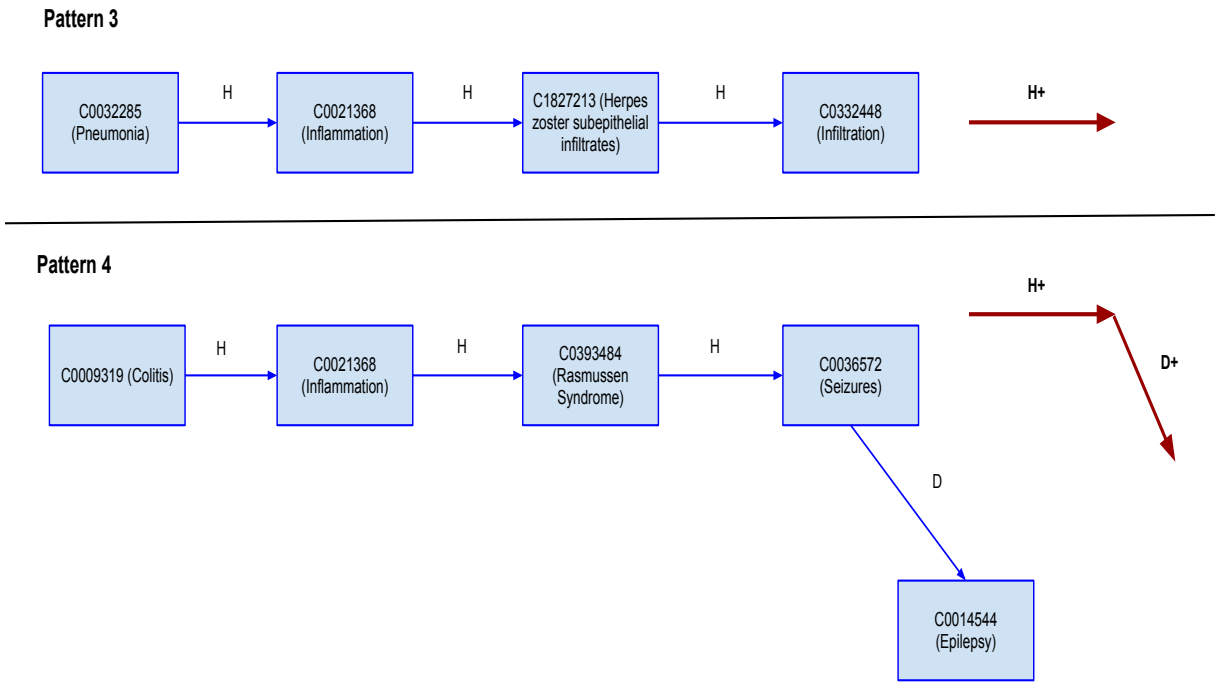

Figure 5.9: Allowable path pattern 3 and 4

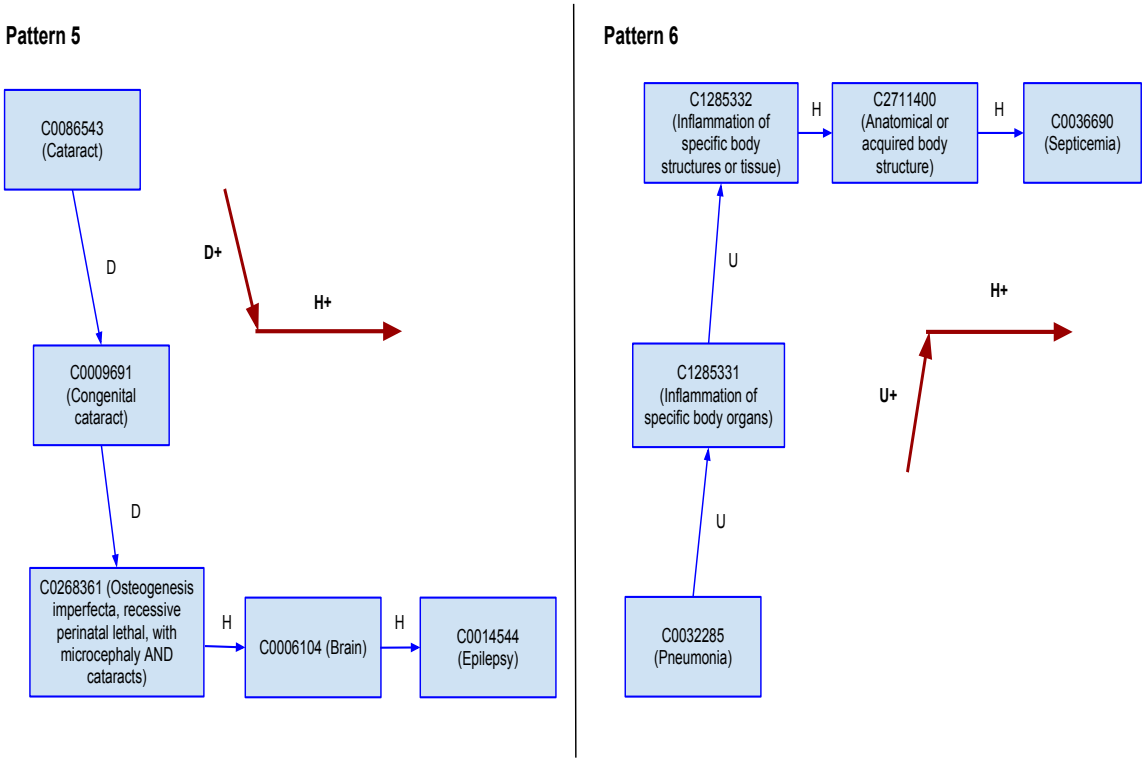

Figure 5.10: Allowable path pattern 5 and 6

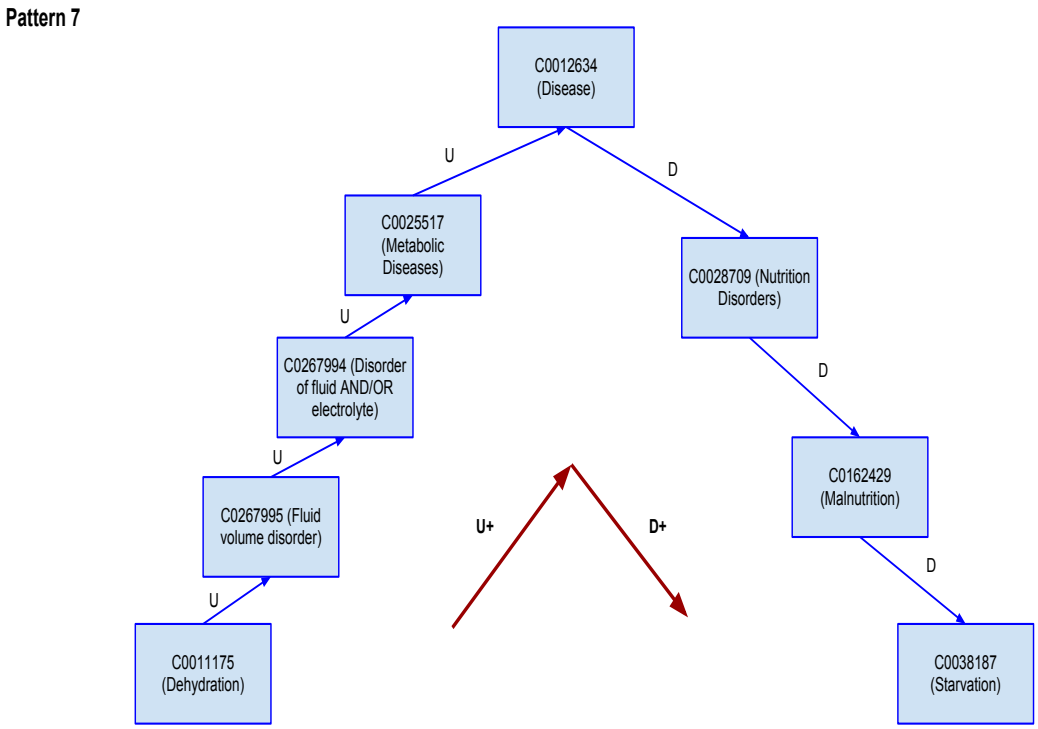

Figure 5.11: Allowable path pattern 7

Pattern 8

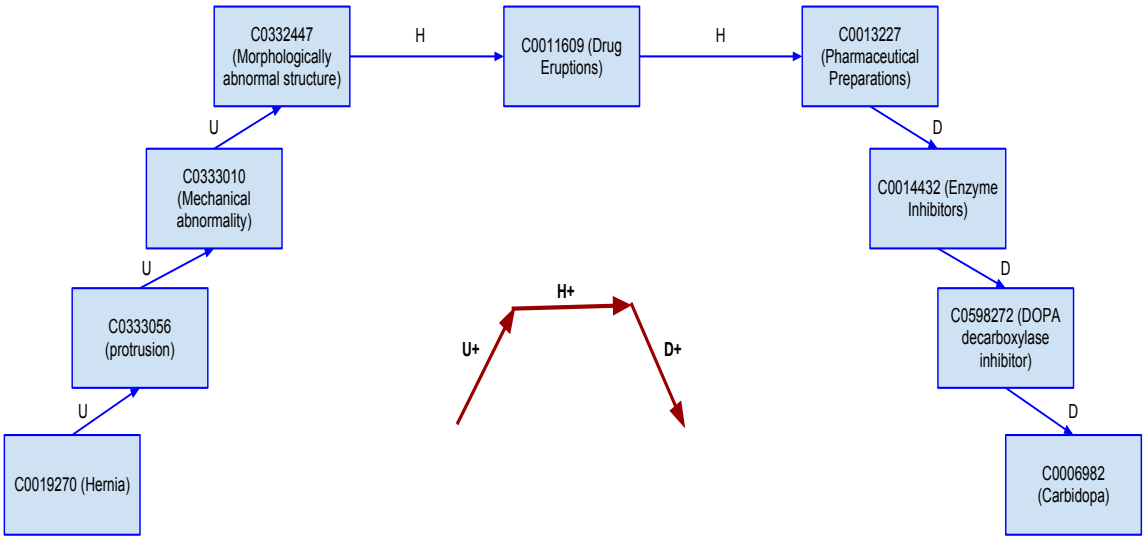

Figure 5.12: Allowable path pattern 8

Thus it is observed that the allowable path patterns suggested by HSO for WordNet, are also found in UMLS SNOMEDCT graph.

Conclusion: Experimental results agree with the hypothesis.

5.6 Hypothesis 6: If the path vectors in an allowable path are restricted in length, it correlates more with gold standard values, as it reduces the number of false positives.

Originally, HSO have not applied any restriction on the allowed length of the vector in either (U, D OR H) direction. The concepts which are not directly connected by a meaningful relation or path, get connected through the root. For example, Cholangiocarcinoma (C0206698) and Colonoscopy (C0009378) are connected through path shown in Figure 5.13.

Cholangiocarcinoma (C0206698) (U) => Adenocarcinoma (C0001418) (U) => Malignant Neoplasms (C0006826) (U) => Neoplastic disease (C1882062) (U) => Neoplasm and/or hamartoma (C1302761) (U) => Disease (C0012634) (U) => Signs and Symptoms (C0037088) (U) => SNOMED CT Concept (SNOMED RT+CTV3) (C2720507) (D) => Interventional procedure (C0184661) (D) => Procedure categorized by device involved (C1285536) (D) => Endoscopy (procedure) (C0014245) (D) => Laparoscopy (C0031150) (D) => Endoscopy of intestine (C0192653) (D) => Endoscopy of large intestine (C0192890) (D) => colonoscopy (C0009378)

Figure 5.13: Path between Cholangiocarcinoma (C0206698) and Colonoscopy (C0009378)

This path of length 14 vector length in upward and downward direction is 7 each. Without restricting the vector length, the CUI pair is assigned semantic relatedness value of 4.5 out of 20, whereas according to gold standards they have a lowest relatedness value in the set. This increases the number of false positives. This section presents experiments performed to see the effect of restricting the path vectors in an allowable path with and without the horizontal relations.

Experiment 8: We configured the HSO algorithm using PAR relation as upward link and CHD relation as downward link for both SNOMEDCT and MSH vocabulary.[8] We then calculated the semantic relatedness values for complete data sets using path vector restriction = 4, 5, 6 and 8. We compare the Spearman's correlation values against

[8] Experiment is performed using Perl package WebService::UMLSKS::Similarity.

baseline correlation values obtained in Experiment 1.

Observations and analysis : To figure out the effect of restricting the vector length, it was necessary to find out correct vector length. Table 5.16 shows the correlation values obtained using default configuration with only up and down vectors and restricting the path vector in each direction to 4, 5, 6 and 8. It also shows correlation values obtained when no restriction is applied on the path vectors, as a point of comparison. The correlation values for all data sets are presented using SNOMEDCT vocabulary. Figure 5.14, 5.15, 5.16 and 5.17 show the path patterns' regular expression, when the path vector is restricted to 4, 5, 6 and 8 respectively.

```
/\bU{1,4}\b|\bU{1,4}D{1,4}\b|\bU{1,4}H{1,4}\b|\bU{1,4}H{1,4}D{1,4}\b
|\bD{1,4}\b|\bD{1,4}H{1,4}\b|\bH{1,4}D{1,4}\b/
```

Figure 5.14: HSO allowable patterns' regular expression after restricting vector length to 4

```
/\bU{1,5}\b|\bU{1,5}D{1,5}\b|\bU{1,5}H{1,5}\b|\bU{1,5}H{1,5}D{1,5}\b
|\bD{1,5}\b|\bD{1,5}H{1,5}\b|\bH{1,5}D{1,5}\b/
```

Figure 5.15: HSO allowable patterns' regular expression after restricting vector length to 5

```
/\bU{1,6}\b|\bU{1,6}D{1,6}\b|\bU{1,6}H{1,6}\b|\bU{1,6}H{1,6}D{1,6}\b
|\bD{1,6}\b|\bD{1,6}H{1,6}\b|\bH{1,6}D{1,6}\b/
```

Figure 5.16: HSO allowable patterns' regular expression after restricting vector length to 6

Table 5.17 shows the correlation values obtained for MiniMayoSRS data set using path restriction of 4, 5, 6 and 8. The results are obtained using MSH vocabulary and are compared against the correlation values obtained with no path restriction.

As it can be observed from both the SNOMEDCT and MSH vocabulary correlation results, restricting the path length to 5, improves the overall correlation with gold

```
/\bU{1,8}\b|\bU{1,8}D{1,8}\b|\bU{1,8}H{1,8}\b|\bU{1,8}H{1,8}D{1,8}\b
|\bD{1,8}\b|\bD{1,8}H{1,8}\b|\bH{1,8}D{1,8}\b/
```

Figure 5.17: HSO allowable patterns' regular expression after restricting vector length to 8

Table 5.16: Spearman's Correlations after restricting vector length (l) to 4, 5, 6 and 8 using default configuration(Figure 5.1)(SAB : SNOMEDCT)

Data set	l = 4	l = 5	l = 6	l = 8	No restriction
MiniMayoSRS.physicians(29)	0.4586	0.5054	0.5532	0.4215	0.3133
	(N = 18)	(N = 20)	(N = 21)	(N = 25)	(N = 26)
MiniMayoSRS.coders(29)	0.7300	0.7619	0.7722	0.6441	0.5102
	(N = 18)	(N = 20)	(N = 21)	(N = 25)	(N = 26)
MayoSRS.gold(101)	0.3977	0.2960	0.2644	0.2109	0.1743
	(N = 41)	(N = 58)	(N = 75)	(N = 86)	(N = 87)
UMNSRS.rel.gold(430)	0.3833	0.3232	0.2857	0.2943	0.2945
	(N = 215)	(N = 282)	(N = 342)	(N = 362)	(N = 363)
UMNSRS.sim.gold(401)	0.5737	0.5128	0.4957	0.5209	0.5247
	(N = 194)	(N = 268)	(N = 318)	(N = 340)	(N = 342)

Table 5.17: Spearman's Correlations after restricting vector length to 4, 5, 6 and 8 using default configuration(Figure 5.2)(SAB : MSH)

Data set	l = 4	l = 5	l = 6	l = 8	No restriction
MiniMayoSRS.msh.physicians(25)	0.3568	0.4978	0.5057	0.5143	0.5143
	(N = 15)	(N = 19)	(N = 21)	(N = 22)	(N = 22)
MiniMayoSRS.msh.coders(25)	0.3975	0.5354	0.5119	0.5403	0.5403
	(N = 15)	(N = 19)	(N = 21)	(N = 22)	(N = 22)

standards. When path vector is restricted to 5, the number of false positives i.e., the CUI pairs which get connected to each other only because they are connected to root of the vocabulary, are decreased. This provides an improvement in the Spearman's correlation coefficient with gold standards.

Experiment 9: We configured the HSO algorithm using PAR relation as upward link, CHD relation as downward link, RB(Broader relation), RN(Narrower relation) and chosen relation attributes set from RO relation by hypothesis 2, as horizontal links.[9] The cost of eachH link is two. We then calculated the semantic relatedness values for big subsets using path vector restriction = 5. We compare the Spearman's correlation values against correlation values obtained in Experiment 4, without the path restriction.

Observations and analysis : We observe the effect of using horizontal relations and path restriction together, for calculating semantic relatedness values. Table 5.18 shows the comparison between the Spearman's correlation values obtained with and without path restriction of 5 and up, down and horizontal vectors are used.

Table 5.18: Comparison between semantic relatedness with and without path restriction after adding H links (SAB : SNOMEDCT)

Data set	No path restriction	vector length = 5
MiniMayo.subset.coders (23)	0.8045	0.7203
	(N = 23)	(N = 23)
MiniMayo.subset.physicians (23)	0.5085	0.4236
	(N = 23)	(N = 23)
Mayo.subset.gold (84)	0.3827	0.3633
	(N = 83)	(N = 70)
UMNSRS.rel.subset.gold (366)	0.3053	0.3244
	(N = 346)	(N = 318)
UMNSRS.sim.subset.gold (341)	0.5128	0.4968
	(N = 323)	(N = 298)

Conclusion: Experimental results support the hypothesis for some of the data sets.

[9] Experiment is performed using Java package WebService::UMLSKS::Similarity::Java.

5.7 Hypothesis 7: Allowing two direction changes in an allowable path between medical concepts, aids in improving the correlation with gold standards.

The original HSO allowable path definition allows only one direction change in the path. After observing the frequently found path patterns in UMLS, we tried to allow two direction changes and analyze the correlation values.

Experiment 10: We configured the HSO algorithm using PAR relation as upward link, CHD relation as downward link, RB(Broader relation), RN(Narrower relation) and chosen relation attributes set from RO relation by hypothesis 2, as horizontal links.[10]
The cost of each H link is two. We then calculated the semantic relatedness values for big subsets (MiniMayoSRS and MayoSRS big subsets) allowing two direction changes. We compare Spearman's correlation values against correlation values obtained in Experiment 4, with only allowing one direction changes as originally suggested by HSO.

Observations and analysis : We observe the effect of allowing two direction changes in an allowable path for calculating semantic relatedness values. The regular expression formed to denote the path patterns after allowing two direction changes is shown Figure 5.18.

```
/\bU+\b|\bU+D+\b|\bD+U+\b|\bU+H+\b|\bU+H+D+\b|\bU+H+U+\b|
\bU+D+H+\b|\bU+D+U+\b|\bD+\b|\bD+H+\b|\bD+H+U+\b|\bD+H+D+\b|
\bH+D+\b|\bH+U+\b|\bH+\b|\bH+D+H+\b|\bH+D+U+\b/
```

Figure 5.18: Patterns with 2 direction changes allowed

The comparison between the correlation values for big subsets (MiniMayoSRS and MayoSRS) is shown in the Table 5.19. The Spearman's correlation coefficient values obtained for both SNOMEDCT and MSH are tabulated to compare the effect of allowing paths with two direction changes, when all U/D and H links are used. These comparisons

[10] Experiment is performed using Java package WebService::UMLSKS::Similarity::Java.

show that when H relations are used along with allowing two direction changes, there is considerable improvement in Spearman's correlation values for 3 out of 5 gold standards. Some paths with two direction changes are shorter than the smallest allowable paths with one direction change. But even if the algorithm finds shorter allowable path when two direction changes are allowed, a higher penalty is applied as number of changes in direction is increased. The shorter paths with higher changes in direction both affect the relatedness values and thus the Spearman's correlation values. For example, Figure **??** shows the variation in the path between C0333997 (Lymphoid hyperplasia) and C0007107 (Malignant neoplasm of larynx). The first path in the figure is obtained when only one direction change allowed and the later path describes a shorter path obtained when two direction changes were allowed in an allowable path.

Path with one direction change allowed:

C0333997(Lymphoid hyperplasia)(H)[RO - has_associated_morphology] -> C1302738(Castleman's superficial pseudotumor, involving skin)(H)[RO - associated_morphology_of] -> C0027651(Neoplasms)(H)[RO - associated_morphology_of] -> C0023055(Laryngeal neoplasm)(D)[CHD - isa] -> C0007107(Malignant neoplasm of larynx)

Semantic Relatedness: 9.75
Path cost: 7
Path Direction: HHHD
Changes in directions: 1

Path with two direction changes allowed:

C0333997(Lymphoid hyperplasia)(U)[PAR - inverse_isa] -> C0024228(Lymphatic Diseases)(D)[CHD - isa] -> C0452190(Malignant immunoproliferative small intestinal disease)(H)[RO - has_associated_morphology] -> C1288351(Malignant Neoplasm (Morphology))(H)[RO - associated_morphology_of] -> C0007107(Malignant neoplasm of larynx)

Semantic Relatedness: 7.0
Path cost: 6
Path Direction: UDHH
Changes in directions: 2

effect

Figure 5.19: Path between C0333997(Lymphoid hyperplasia) and C0007107(Malignant neoplasm of larynx)

Table 5.19: Comparison between correlation values with one direction (1D) and two direction (2D) changes allowed using H relations (SAB : SNOMEDCT and MSH)

Data set	1D change (SNOMEDCT)	2D change (SNOMEDCT)	1D change (MSH)	2D change (MSH)
MiniMayo.subset.coders (23)	0.8045 (N = 23)	0.7726 (N = 23)	0.6545 (N = 25)	0.6693 (N = 22)
MiniMayo.subset.physicians (23)	0.5085 (N = 23)	0.4296 (N = 23)	0.6249 (N = 22)	0.6369 (N = 22)
Mayo.subset.gold (84)	0.3827 (N = 83)	0.4144 (N = 83)	NA	

Conclusion: Experimental results does support the hypothesis.

Summary of Results: To summarize the results from various experiments performed, we can conclude that HSO's application on UMLS graph using SNOMEDCT and MSH vocabulary, led to interesting findings and also shows the areas where HSO's calculation of semantic relatedness can be improved for medical terms. It is seen that when the HSO measure is applied with only up (PAR relation) and down (CHD relation) vectors, it can be defined equivalently to the shortest path measure implemented by UMLS::Similarity package. Further we confirm that it is necessary to filter the pool of horizontal relations and attributes and find the useful attributes that can be used for the calculation of relatedness. When these useful relations and attributes are used as horizontal links with up and down links, we improve the correlation to the gold standards for medical concepts. In the later experiments we confirm that that the cost of a horizontal link should be greater than a up or down link, as suggested by HSO algorithm. We then explore the path patterns found in SNOMEDCT vocabulary and find that all allowable path patterns from HSO's allowable pattern set can be observed in UMLS's SNOMEDCT vocabulary graph. Furthermore, we show that the Spearman's correlation coefficient value can be further improved by restricting the allowable path vectors in length. Finally we observe that improvement in correlation values can be achieved by allowing two direction changes in an allowable path.

Chapter 6

Related Work

There is considerable work done in past to find the semantic relatedness and semantic similarity between the medical terminologies. A variety of Ontology-based approaches including path-based, node-based, feature-based and combination approaches have been tried to find the semantic relatedness in biomedical domain. Some of the prominent approaches are described in brief in this chapter.

6.1 Application of HSO for Malapropism Detection

Hirst and St-Onge(HSO) [6] applied their measure of semantic relatedness to detect and correct the malapropisms in the text using WordNet vocabulary graph in 1998. A malapropism is the confusing substitution of an intended word with another word of similar sound or similar spelling that has a quite inappropriate and different meaning, e.g., Success is the defect(effect) of hard work.

HSO measure also present a way of using cohesive relations for constructing Lexical or cohesive Chains [6]. These chains are used by HSO measure in one of its application to detect malapropisms. In a coherent and cohesive texts, it is observed that successive sentences tend to use words related to concepts used in previous sentences. The words which share the similar concept can be thought of forming a lexical chain. A Lexical chain is a sequence of semantically related words in the text that provides the information about the context of the text. This chain is independent of the grammatical structure of the text and may span short(nearby words) or long distances(complete

text). There may be multiple lexical chains in the same text.
To illustrate this, consider following example:

Rome is the capital of Greece. There are lot of resources on web about the inhabitants of the city. On-line information helps to update knowledge about new places.

This text has two lexical chains:

- Rome - capital - Greece - inhabitants - city - places
- resources - web - On-line - information - knowledge

These two lexical chains describe two different concepts and words in them are related to each other by a particular cohesive relation. The first one tells that the text is about Rome, Greece and its inhabitants. The second lexical chain tells that text is about resources and information on web. Thus complete context of the text can be understood using these two lexical chains.

As lexical chains connect semantically related words together and express the context of the text, they are used to detect a malapropism, a word that does not fit into the context. Any word that can not be inserted into any of the lexical chains formed for the surrounding text is assumed to a potential malapropism. HSO's application for malapropism detection provides a detail algorithm for the formation of lexical chains using WordNet thesaurus, followed by detection and correction of malapropism. The work also presents the analysis of results which show that HSO measure can be successfully used for detection and correction of malapropisms.

6.2 Development of a conceptual distance metric for the UMLS

Jorge E. Caviedes and James J. Cimino presented a conceptual distance metric for UMLS framework in 2004 [23]. Along with explaining the general conceptual and lexical matching principles, they describe the algorithm used for calculating conceptual metrices between medical concepts and between set of concepts (i.e., cluster of concepts). Conceptual matching is useful in finding the similarity score between two concepts and

thus provides a way of using similar concepts in absence of exact lexical match. Calculating conceptually similar concepts becomes important for computer applications that perform approximate matching, inferencing and data mining biomedical informatics.

Caviedes and Cimino propose a method of quantifying semantic similarity between medical concepts based on minimum number of parent links between them. Conceptual matching metric assigns a semantic similarity score between two concepts, such that identical concepts are assigned 0 score and dis-similar concepts get higher score value. As, the conceptual distance between concepts increases, the score also increases. They use three sets of concepts for source vocabularies/terminologies, viz., MeSH, ICD9CM, and SNOMED from UMLS to evaluate the performance of the algorithm. The evaluation is done by comparing the results against human judgments.

The algorithm targets to develop metrices for concepts within single terminology, from multiple chosen terminologies and without specifying any vocabulary. To cover the concepts from different vocabulary the algorithm uses PAR(IS-A relation) links and RB (Broader relation) links to find shortest path between concepts. The distance of the shortest path without cycles in subgraph formed by using only PAR and RB links, is termed as the conceptual distance between two concepts. These conceptual distances were then compared against the expert scores provided by physicians.

The results of calculating conceptual distances for concepts from three vocabularies (MSH, SNMI, ICD9CM), show that conceptual distances in MSH show highest correlation with expert scores. They also perform experiments for finding conceptual distance between unrelated concepts and concepts in the clusters. Finally it is concluded that conceptual distance metric can be used as an indication of similarity between concepts and between cluster of concepts. The metric can be used for one or more source vocabularies.

6.3 Measures of semantic similarity and relatedness in the biomedical domain

In 2006, Pedersen, et. al.[20], applied six measures of semantic similarity and relatedness from independent domains on UMLS graph structure. They adopted these measures to perform domain-specific tasks of biomedical domain. The measures were developed

originally for WordNet, a English vocabulary database. SNOMEDCT source vocabulary was chosen to perform the experiments. Such measures can be very useful in information retrieval, document retrieval, indexing for medical data and spelling correction.

The six measures include:

- Two Path based measures
- three measures that use path information with information content statistics
- a measure based on context vectors

Thus the work presents comparison between the different measures of semantic relatedness and similarity, such as : Path Length, Wu & Palmer, Leacock & Chodorow, Hirst & St-Onge, Resnik, Jiang & Conrath; Lin and Patwardhan & Pedersen, while presenting the advantages and disadvantages of each of them.

To apply these six domain-independent measures to medical domain they chose three knowledge sources from the domain, viz., SNOMEDCT (Systematized Nomenclature of Medicine, Clinical Terms), a major terminology in UMLS, Mayo clinic corpus of clinical notes and Mayo clinic thesaurus. The six measures were evaluated against a test set of 30 medical concept pairs newly created by Pedersen, et. al. This test set was scored by coders and physicians, resulting to two gold standards. A Spearman's rank correlation coefficient was used to evaluate each measure's performance for the test data when compared against human judgments. It was concluded that the domain-independent measures can be adopted to a specific medical domain effectively.

6.4 Comparison of Ontology-based Semantic- Similarity Measures

In 2008, Lee et.el[24] developed a semantic-similarity measure which calculates the concept similarity. The similarity values are calculated within an ontology. An ontology is a declarative model that defines and represents the concepts existing in a domain. It also defines concept attributes and the relationships between them.[1] This work compares three semantic-similarity approaches applied to SNOMEDCT. The comparison

[1] http://www.openclinical.org/ontologies.html

was done by comparing semantic-matrices based on expert opinion, ontologies only and information content.

The system uses a prototype system for ontology-based annotation of resource elements, which is one of the tools offered by BioPortal at the National Center for Biomedical Ontology (NCBO). A test sample of 20 diseases was shortlisted by a trained physician from the broad range of 225 disease concepts mainly focused by Genomic Nosology for Medicine (GNOMED) project. Expert assessments, Cluster-Based metric, Term-Frequency (TF) Metric, and Descendant-Distance (DD) Metric were applied to 190 distinct pairwise combinations of the test sample disease concepts. The comparison results showed that metrics based on information content poorly agree with the ontology only metric. Whereas, the ontology only metric correlated most with expert opinion.

6.5 UMLS::Similarity

There are different semantic similarity measures developed using different medical sources which makes it difficult to implement them consistently and compare their results. To solve these problem, Pedersen, et al.[25] developed two frameworks named UMLS::Interface and UMLS::Similarity which allow the developers to test their measure and compare the results with other available measures in 2009. There are two other packages named UMLS Knowledge Source(UMLSKS) and UMLS Query which allow the programmers to interact with UMLS. UMLS::Similarity is a package which obtains semantic similarity between terms when their sources and relations are specified. It can be used to test the newly developed measure and then the results can be compared with other measures.

UMLS-Similarity contains five previously developed semantic similarity measures proposed by Rada, et. al., Wu and Palmer, Leacock and Chodorow, and Nguyen and Al-Mubaid, and the Path measure. It uses 29 term pairs to evaluate the semantic measures and non-parametric Spearman's rank correlation coefficient to compare the results with values provided by coders and physicians. Thus UMLS-Similarity is used to evaluate HSO measure presented by this thesis, as it is reliable framework to test the semantic similarity measures developed for UMLS[25].

Chapter 7

Conclusion

The thesis starts by applying the HSO formulation along with the concept of an allowable path on UMLS graph structure. Experimental results confirm the fact that the structure of UMLS is different from that of WordNet. UMLS is bigger and denser in it's structure leading to interesting set of paths between the concepts. We observe that HSO's suggestion of *penalizing an allowable path for each direction change it makes, leads to accurate calculation of semantic relatedness*, holds true even for medical concepts.

It is seen that the correlation obtained by using the HSO measure with upward and downward links is comparable to the correlation values obtained by using path measure with UMLS-Similarity. Further we confirm that it is necessary to filter the pool of horizontal relations and attributes and find the useful attributes that can be used for the calculation of relatedness.

On the basis of the experiments performed with the selected relations and attributes, the results shows that including horizontal links leads to better correlation to gold standards than using only upward and downward links. To perform the experiments with large set of selected horizontal relations and attributes, we filter the concepts that are connected to unmanageable number of neighboring concepts and come up with big subsets of the experimental data sets.

In the later experiments we prove that that the cost of a horizontal link should be greater than a up or down link, as suggested by the HSO measure. We then explore the path patterns found in SNOMEDCT vocabulary and find that all allowable path

patterns from HSO's allowable pattern set can be observed in UMLS's SNOMEDCT vocabulary graph.

We find that restricting the path vector length to 5 in each direction of path reduces the number of false positives, i.e., some non-related concepts which get connected by a longer path mostly going through root of the vocabulary. This improves the correlation value further, resulting in improvement in correlation values with gold standards. After observing that the path patterns present in UMLS are different than those of WordNet, the thesis tries to experiment with patterns other than the original HSO allowable paths patterns. It allows two direction changes in an allowable path as compared to one, in the HSO allowable pattern set. This experiment shows that semantic relatedness values correlate more to gold standards than allowing one direction change. Thus the thesis makes an effort to improve the HSO measure and accommodate the structure and relations of UMLS.

Thus we can conclude that the HSO measure can be efficiently extended by including following:

- Penalizing an allowable path exponentially for each direction change it makes.
- Including wisely chosen set of relations and attributes as horizontal links.
- Restricting the vector length in each direction of an allowable path.
- Allowing two direction changes in an allowable path.

Chapter 8

Future Work

While calculating the semantic relatedness between medical concepts using UMLS, experimental results confirmed a need for choosing the right set of relations and relation attributes that form the upward, downward and horizontal links. This thesis work presents a configuration of relations and attributes that leads to improvement in the performance of measure. We feel that there is a further room for exploring and experimenting with more relations and attributes. We see improvement in the Spearman's correlation values after filtering out temporal and qualifying relation attributes such as 'severity of', 'episodacity of','was-a', etc. The set of relations can further be refined in future by consulting with doctors and clinical experts.

Currently UMLSKS APIs return unfiltered and detailed response for each query made to the server. UMLSKS API does not support functions that return specific information about a CUI/term. This results in extra overhead of parsing the response and extracting the required information on the client machine. This also increases the algorithm's space and time complexity, limiting the amount of concepts that can be handled by system at a time. Currently both PERL and JAVA implementations of the algorithm suffer due to this overhead and can further be improved by using updated UMLSKS API in future. UMLS plans to provide updated granular web services in future that would allow users to request for specific information about the concept. These web services can be used to make the system more efficient, simpler and thus more scalable.

Presently, the system is also affected by network bottlenecks resulting in variable performance. We analyzed the time taken by different parts of the system along with

the time taken for querying UMLS. The timing analysis showed that approximately 60% of the total time taken by program was spent in accessing UMLS database through web services. As discussed earlier, some problem CUIs from UMLS have unmanageable number(greater than 1000) of concepts connected to them. Accessing such problem CUIs with all their neighbors over the network makes the computations space and time consuming. This problem can be solved in future by choosing only useful neighboring concepts for the source concept. This can achieved by seeking doctors' opinion and using UMLSKS improved web services and developers' support in future.

In this thesis we experiment with HSO's original allowable path set and modified path sets formed by restricting the path vector length in each direction and by allowing two direction changes in the path. Experimental results show that including these new allowable path patterns for UMLS, leads to interesting results. As UMLS graph structure is denser and larger than that of WordNet, there are various path patterns that can be good candidates for an allowable path. Work can be done in this regard to find more such patterns and include them in an allowable path set which will improve the HSO measure for UMLS.

We mainly used SNOMEDCT as a primary source vocabulary and MSH vocabulary for calculating the semantic relatedness for experimental data. In future other source vocabularies such as NCBI, FMA, GO etc., can be used to explore new relationships between concepts. It would also be exciting to perform experiments combining different source vocabularies which would let system find relatedness between inter-vocabulary concepts.

References

[1] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270, September 2004.

[2] A. Budanitsky. Lexical Semantic Relatedness and Its Application in Natural Language Processing. `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.34.1036`, August 1999.

[3] C. E. Osgood. The nature and measurement of meaning. *Psychological Bulletin*, 49(3):197–237, May 1952.

[4] M. R. Quillian. Semantic memory. In M. Minsky, editor, *Semantic Information Processing*, pages 227–270. MIT Press, 1968.

[5] A. M. Collins and E. F. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428, November 1975.

[6] G. Hirst and D. St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In Fellbaum C., editor, *WordNet:An electronic Lexical Database*. The MIT Press, Cambridge, 1998.

[7] G. A. Miller. WordNet: A Lexical Database for English. *COMMUNICATIONS OF ACM*, 38(11):39–41, November 1995.

[8] National Library of Medicine(US). *UMLS Reference Manual*, September 2009.

[9] National Library of Medicine(US). Umls Source Release Documentation. `http://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html`, November 2008.

[10] UMLS Terminology Services. SNOMED CT Browser. https://uts.nlm.nih.gov//snomedctBrowser.html.

[11] Kent Spackman. SNOMED Clinical Terms Fundamentals. http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/SNOMED_Clinical_Terms_Fundamentals.pdf, December 2007.

[12] International Health Terminology Standards Development. SNOMED CT User Guide. http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_UserGuide_Current-en-US_INT_20120131.pdf, January 2012.

[13] Stuart Nelson, M.D. Fact Sheet Medical Subject Headings (MeSH). http://www.nlm.nih.gov/pubs/factsheets/mesh.html, December 1999.

[14] National Library of Medicine(US). Unified Medical Language System (UMLS) Basics Tutorial. http://www.nlm.nih.gov/research/umls/new_users/online_learning/index.htm, October 2008.

[15] Paul Kulchenko. SOAP:Lite for Perl - SOAP Cookbook. http://cookbook.soaplite.com/, 2001.

[16] UMLS Terminology Services. UMLS Technology Services- Developer's Guide. https://uts.nlm.nih.gov//doc/devGuide/index.html, 2011.

[17] H. Rubenstein and J. Goodenough. Contextual correlates of synonymy. *COMMUNICATIONS OF ACM*, 8:627–33, 1965.

[18] S. Patwardhan, S. Banerjee, and T. Pedersen. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the fourth international conference on intelligent text processing and computational linguistics*, Mexico City, Mexico, 2003.

[19] University of Minnesota. University of Minnesota Pharmacy Informatics Lab. http://rxinformatics.umn.edu/SemanticRelatednessResources.html.

[20] T. Pedersen and S. Patwardhan and S.V.S. Pakhomov and C.G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, 2007.

[21] S. Pakhomov, B. McInnes, T. Adams, Y. Liu, G.B. Melton, and T. Pedersen. Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, Washington D.C., November 2010.

[22] `http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient#cite_note-myers2003-0`.

[23] J. E. Caviedes and J. J. Cimino. Towards the development of a conceptual distance metric for the UMLS. *Journal of Biomedical Informatics*, 37:77–85, 2004.

[24] W. Lee, N. Shah, K. Sundlass, and M. Musen. Comparison of Ontology-based Semantic-Similarity Measures. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*, Published Online, 2008.

[25] B.T. McInnes, T. Pedersen, and S.V. Pakhomov. UMLS-Interface and UMLS-Similarity: Open source software for measuring paths and semantic similarity. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*, San Francisco, CA, 2009.

Appendix A

Additional Results

This section provides the additional experimental results with different configurations for MayoSRS data set. The semantic relatedness values for all the CUI pairs MayoSRS data set are tabulated for different experiments.

Table A.1: MayoSRS test set (part 1)

Term 1(CUI 1)	Term 2(CUI 2)	SR
Metastasis, Neoplasm(C0027627)	Carcinomatoses(C0205699)	8.23
Rheumatoid arthritis NOS(C0003873)	Nodule, Rheumatoid(C0035450)	7.08
Stomatitis NOS(C0038362)	Ulcer, Oral(C0149745)	6.85
Ileitis, NOS(C0020877)	Crohn's disease NOS(C0010346)	6.85
Walking difficulties(C0311394)	Antalgic gait(C0231685)	6.69
Rheumatoid arthritis NOS(C0003873)	(C1396859)	6.15
Confusion(C0009676)	Delusion, NOS(C0011253)	6.08
Hallucinations NOS(C0018524)	Disorders, Psychotic(C0033975)	6
joint morning stiffness(C0457086)	Rheumatoid arthritis NOS(C0003873)	5.69
Stenosis, Aortic Valve(C0003507)	Calcification, Physiologic(C0006660)	5.62
Diarrhea NOS(C0011991)	COLITIS (NOS)(C0009319)	5.54
Hemolysis (disorder)(C0019054)	Haemoglobin finding(C1561562)	5.46
Colonoscopy, NOS(C0009378)	Polyp, NOS(C0032584)	5.46
Syndromes, Paraneoplastic(C0030472)	Malignantneoplasm, primary(C1306459)	5.31
Urticaria NOS(C0042109)	Butterfly rash(C0277942)	5.31
Pain, Back(C0004604)	Stenosis, Spinal(C0037944)	5.31
T wave feature(C0429103)	Infarction, Myocardial(C0027051)	5.23
Antinuclear Antibody(C0003243)	autoimmune reactions(C0443146)	5.23
Dyspepsia, NOS(C0013395)	Ulcer, Peptic(C0030920)	5.23
leukaemias(C0023418)	cells stems(C0038250)	5.08
Scleroderma NOS(C0011644)	Scleroderma, Systemic(C0036421)	5
Cerebrovascular accident, NOS(C0038454)	Hemipareses(C0018989)	5
Pneumoniae(C0032285)	Infiltration, NOS(C0332448)	4.85
phenomenon raynauds(C0034735)	Ischaemia, NOS(C0022116)	4.85
Calculi, Kidney(C0022650)	Ureteral Obstructions(C0041956)	4.69
Temporal Arteritides(C1956391)	Headache, NOS(C0018681)	4.69
Myopathies(C0026848)	Dermatomyositides(C0011633)	4.46
Walking (activity)(C0080331)	climbing stair(C0432601)	4.38
Deglutition, NOS(C0011167)	Peristalsis, NOS(C0031133)	4.38
Gastrostomy, NOS(C0017196)	Malnutrition NOS(C0162429)	4.31
Brain Stems(C0006121)	Entire cranial nerve(C1269897)	4.23
Dyspnoea, NOS(C0013404)	Tachypnoea(C0231835)	4.08
Cavitation, NOS(C1510420)	Tuberculosis NOS(C0041296)	4.08

Table A.2: MayoSRS test set (part 2)

Term 1(CUI 1)	Term 2(CUI 2)	SR
Asthenia NOS(C0004093)	neuropathies(C0442874)	4.08
Osteophytes(C1956089)	heberdens nodes(C0018862)	4
Ulcer, Leg(C0023223)	Varicosities(C0042345)	3.92
Peripheral oedema(C0085649)	Oedema - pulmonary NOS(C0034063)	3.92
drawers sign(C0231736)	pain in knee(C0231749)	3.77
Rheumatoid arthritis NOS(C0003873)	Arthroscopy, NOS(C0003904)	3.77
Disorders, Deglutition(C0011168)	hypomotility(C0679317)	3.77
points trigger(C0458343)	Fibromyalgia, NOS(C0016053)	3.54
Seizure NOS(C0036572)	Headache, NOS(C0018681)	3.31
Arrhythmia, Cardiac(C0003811)	Valve, Mitral(C0026264)	3.31
Injection (procedure)(C1533685)	Hydrarthrosis, NOS(C1253936)	3.31
Pulmonary Embolisms(C0034065)	Pneumoniae(C0032285)	3.23
Panniculitis NOS(C0030326)	Lipoma, NOS(C0023798)	3.15
Malaria NOS(C0024530)	Amoebiasis NOS(C0002438)	3.15
Vasculitis, NOS(C0042384)	Thrombosis, NOS(C0040053)	3
Penicillin(C0030842)	Hypersensitivity NOS(C0020517)	3
Degenerative polyarthritis,NOS(C0029408)	Bony sclerosis(C0221434)	2.85
Sinusitis NOS(C0037199)	Sinusoidal(C0442041)	2.85
cortisones(C0010137)	Family history: Osteoporosis(C1563292)	2.85
neuropathies(C0442874)	paralyse(C0522224)	2.85
Pulmonary Embolisms(C0034065)	Hemoptysis NOS(C0019079)	2.77
Meniscus structure of joint(C0224498)	Degenerative polyarthritis,NOS(C0029408)	2.62
Nodule, Rheumatoid(C0035450)	PULMONARY NODULE(C0034079)	2.38
oedemas(C0013604)	Rate, Glomerular Filtration(C0017654)	2.38
Dyspareunia (female)(C0013394)	Ovulations(C0029965)	2.31
(C2267026)	hyperlipidaemias(C0020473)	2.23
Uveitis NOS(C0042164)	Antigen, HLA-B27(C0019740)	2.23
immunisations(C0020971)	Syndromes, Immunologic Deficiency(C0021051)	2.15
Laxity, NOS(C0332536)	Syndrome, Marfan(C0024796)	2.08
corneal ulcers(C0010043)	Ulcer, Pressure(C0011127)	2
Dysgeusias(C0013378)	Deglutition, NOS(C0011167)	1.92
Necrosis, NOS(C0027540)	*Liquefaction(C1547030)	1.92
Prothrombin(C0033706)	Syringe, NOS(C0039142)	1.77
Malignant Prostate Neoplasm(C0376358)	Acid Phosphatase(C0001109)	1.69

Table A.3: MayoSRS test set (part 3)

Term 1(CUI 1)	Term 2(CUI 2)	SR
Systemic infections(C0243026)	Hypotension NOS(C0020649)	1.69
congestive failures heart(C0018802)	Portal Hypertensions(C0020541)	1.69
Erythema NOS(C0041834)	Osteoporosis NOS(C0029456)	1.54
HAEMATEMESIS(C0018926)	Xerostomia(C0043352)	1.54
Sarcoidosis, NOS(C0036202)	Vitamin D, NOS(C0042866)	1.46
Aneurysm, Cerebral(C0917996)	Pulmonary Embolisms(C0034065)	1.46
Perseveration(C0233651)	Ulcers, Venous Stasis(C1527356)	1.31
Ketoacidosis, NOS(C0220982)	Lupus erythematosus NOS(C0409974)	1.23
Antibodies, Antiphospholipid(C0162595)	Acne NOS(C0702166)	1.08
Ligament rupture(C0262538)	Pointes, Torsades de(C0040479)	1.08
Chronic Obstructive Disease(C0024117)	Halitoses(C0018520)	1.08
Myeloma-Multiple(C0026764)	Depressive disorder NOS(C0011581)	1.08
Fibrosis, Pulmonary(C0034069)	Appendicitis NOS(C0003615)	1
Diabetes mellitus NOS(C0011849)	Nodule, Rheumatoid(C0035450)	1
Portal Hypertensions(C0020541)	melanocytic naevi(C0027962)	1
Lymphoid hyperplasia, NOS(C0333997)	Xerostomia(C0043352)	1
heart failures(C0018801)	Disorders, Psychotic(C0033975)	1
cirrhosis cryptogenic(C0267809)	gastrins(C0376180)	1
Diabetes mellitus NOS(C0011849)	Polyp, NOS(C0032584)	1
Small Cell Carcinoma(C0149925)	DiabetesMellitus(C0011860)	1
Stenosis, Mitral Valve(C0026269)	Ulcer, Peptic(C0030920)	1
Agents, Contraceptive(C0009871)	Contraindicated(C1444657)	1
Adult respiratory syndrome(C0035222)	Cellulitis, NOS(C0007642)	1
depression bipolar(C0005587)	Atrioventricularjunctional (C0232208)	1
Stenosis, Tracheal(C0040583)	Malignant Tumor (C0007102)	1
Sodium (NOS)(C0037473)	Imaging, Magnetic Resonance(C0024485)	1
Dyspnoea, NOS(C0013404)	Myalgia NOS(C0231528)	1
compression spinal cord(C0037926)	caring wound(C0886052)	1
Syndrome, Carpal Tunnel(C0007286)	Alveolitis(C0549493)	1
splinting hand(C0409162)	splinter hemorrhages(C0333286)	1
rectal polyps(C0034887)	Neoplasms, Laryngeal(C0023055)	1
Hepatitides, Autoimmune(C0241910)	Obstetric Labor, Premature(C0022876)	1
Spinal stenosis of cervical region(C0158280)	liver cirrhosis(C0023891)	1
Degenerative polyarthritis,NOS(C0029408)	Nephrolithiasis NOS(C0392525)	1

Table A.4: MiniMayoSRS set (SAB:SNOMEDCT, REL:PAR, DIR:U)

Source CUI	Destination CUI	SR
Renal failure (C0035078)	Kidney failure (C0035078)	20.00
Abortion (C0156543)	Miscarriage (C0000786)	19.00
Heart (C0018787)	Myocardium (C0027061)	17.00
Metastasis (C0027627)	Adenocarcinoma (C0001418)	13.50
Pulmonary brosis (C0034069)	Lung cancer (C0242379)	12.75
Brain tumor (C0006118)	Intracranial hemorrhage (C0151699)	12.75
Antibiotic (C0003232)	Allergy (C0020517)	11.25
Depression (C0011581)	Cellulitis (C0007642)	11.25
Multiple sclerosis (C0026769)	Psychosis (C0033975)	11.25
Congestive heart failure (C0018802)	Pulmonary edema (C0034063)	10.50
Diarrhea (C0011991)	Stomach cramps (C0344375)	10.50
Mitral stenosis (C0026269)	Atrial brillation (C0004238)	10.50
Pulmonary embolus (C0034065)	Myocardial infarction (C0027051)	10.50
Rheumatoid arthritis (C0003873)	Lupus (C0409974)	10.50
Carpal tunnel syndrome (C0007286)	Osteoarthritis (C0029408)	9.75
Lymphoid hyperplasia (C0333997)	Laryngeal cancer (C0007107)	9.75
Diabetes mellitus (C0011849)	Hypertension (C0020538)	9.75
Hyperlipidemia (C0020473)	Metastasis (C0027627)	9.00
Xerostomia (C0043352)	Alcoholic cirrhosis (C0023891)	9.00
Appendicitis (C0003615)	Osteoporosis (C0029456)	9.00
Peptic ulcer disease (C0030920)	Myopia (C0027092)	8.25
Cortisone (C0010137)	Total knee replacement (C0086511)	7.50
Rectal polyp (C0034887)	Aorta (C0003483)	6.00
Acne (C0702166)	Syringe (C0039142)	6.00
Stroke (C0038454)	Infarct (C0021308)	5.25
Varicose vein (C0042345)	Entire knee meniscus (C0224701)	5.25
Delusion (C0011253)	Schizophrenia (C0036341)	4.50
Cholangiocarcinoma (C0206698)	Colonoscopy (C0009378)	4.50
Calcication (C0175895)	Stenosis (C0009814)	-1.00

Table A.5: MayoSRS set 1 (SAB : SNOMEDCT, REL : PAR, DIR : U)

Source CUI	Destination CUI	SR
Metastasis, Neoplasm(C0027627)	Carcinomatoses(C0205699)	18
Myopathies(C0026848)	Dermatomyositides(C0011633)	17
Vasculitis, NOS(C0042384)	Thrombosis, NOS(C0040053)	12.75
drawers sign(C0231736)	pain in knee(C0231749)	12.75
Peripheral oedema(C0085649)	Oedema - (C0034063) pulmonary NOS	12.75
Deglutition, NOS(C0011167)	Peristalsis, NOS(C0031133)	12.75
corneal ulcers(C0010043)	Ulcer, Pressure(C0011127)	12.75
Malaria NOS(C0024530)	Amoebiasis NOS(C0002438)	12.75
Pulmonary Embolisms(C0034065)	Hemoptysis NOS(C0019079)	12.00
Ulcer, Leg(C0023223)	Varicosities(C0042345)	12.00
Calculi, Kidney(C0022650)	Ureteral Obstructions(C0041956)	12.00
Temporal Arteritides(C1956391)	Headache, NOS(C0018681)	12.00
Stomatitis NOS(C0038362)	Ulcer, Oral(C0149745)	12.00
Ileitis, NOS(C0020877)	Crohn's disease NOS(C0010346)	12.00
Seizure NOS(C0036572)	Headache, NOS(C0018681)	11.25
Pulmonary Embolisms(C0034065)	Pneumoniae(C0032285)	11.25
Syndromes,(C0030472) Paraneoplastic	Malignantneoplasm,(C1306459) primary	11.25
Hemolysis(C0019054) (disorder)	Haemoglobin finding(C1561562)	11.25
Urticaria NOS(C0042109)	Butterfly rash(C0277942)	11.25
Diabetes mellitus NOS(C0011849)	Polyp, NOS(C0032584)	11.25
Dyspnoea, NOS(C0013404)	Tachypnoea(C0231835)	11.25
Rheumatoid arthritis NOS(C0003873)	Nodule,(C0035450) Rheumatoid	11.25
Adult respiratory (C0035222) distress syndrome	Cellulitis, NOS(C0007642)	11.25
Pain, Back(C0004604)	Stenosis, Spinal(C0037944)	11.25
congestive failures (C0018802) heart	Portal(C0020541) Hypertensions	11.25
Diabetes mellitus NOS(C0011849)	Nodule,(C0035450) Rheumatoid	10.50
joint morning(C0457086) stiffness	Rheumatoid arthritis NOS(C0003873)	10.50
Agents, Contraceptive(C0009871)	Contraindicated(C1444657)	10.50
Dyspnoea, NOS(C0013404)	Myalgia NOS(C0231528)	10.50
Dyspepsia, NOS(C0013395)	Ulcer, Peptic(C0030920)	10.50
Fibrosis, Pulmonary(C0034069)	Appendicitis NOS(C0003615)	9.75
Perseveration(C0233651)	Ulcers, Venous Stasis(C1527356)	9.75
Erythema NOS(C0041834)	Osteoporosis NOS(C0029456)	9.75

Table A.6: MayoSRS set 2 (SAB : SNOMEDCT, REL : PAR, DIR : U)

Source CUI	Destination CUI	SR
Systemic infections(C0243026)	Hypotension NOS(C0020649)	9.75
oedemas(C0013604)	Rate, Glomerular Filtration(C0017654)	9.75
Panniculitis NOS(C0030326)	Lipoma, NOS(C0023798)	9.75
Cerebrovascular(C0038454) accident, NOS	Hemipareses(C0018989)	9.75
Brain Stems(C0006121)	Entire cranial nerve(C1269897)	9.75
Diarrhea NOS(C0011991)	COLITIS (NOS)(C0009319)	9.75
Spinal stenosis (C0158280) of cervical region	of alcoholic (C0023891) liver cirrhosis	9.75
points trigger(C0458343)	Fibromyalgia, NOS(C0016053)	9.00
Osteophytes(C1956089)	heberdens nodes(C0018862)	9.00
Sarcoidosis, NOS(C0036202)	Vitamin D, NOS(C0042866)	9.00
immunisations(C0020971)	Syndromes, Immunologic (C0021051) Deficiency	9.00
Portal Hypertensions(C0020541)	melanocytic naevi(C0027962)	9.00
Laxity, NOS(C0332536)	Syndrome, Marfan(C0024796)	9.00
Lymphoid hyperplasia, NOS(C0333997)	Xerostomia(C0043352)	9.00
heart failures(C0018801)	Disorders, Psychotic(C0033975)	9.00
Penicillin(C0030842)	Hypersensitivity NOS(C0020517)	9.00
Chronic Obstructive(C0024117) Airways Disease	Halitoses(C0018520)	9.00
Dysgeusias(C0013378)	Deglutition, NOS(C0011167)	9.00
rectal polyps(C0034887)	Neoplasms, Laryngeal(C0023055)	9.00
HAEMATEMESIS(C0018926)	Xerostomia(C0043352)	9.00
phenomenon raynauds(C0034735)	Ischaemia, NOS(C0022116)	8.25
Gastrostomy, NOS(C0017196)	Malnutrition NOS(C0162429)	8.25
Walking (activity)(C0080331)	climbing stair(C0432601)	8.25
Small Cell Carcinoma (C0149925) of the Lung	DiabetesMellitus(C0011860)	8.25
Cavitation, NOS(C1510420)	Tuberculosis NOS(C0041296)	8.25
Myeloma-Multiple(C0026764)	Depressive disorder NOS(C0011581)	8.25
Stenosis, Tracheal(C0040583)	Malignant Tumor (C0007102) of the Colon	8.25
Sodium (NOS)(C0037473)	Imaging, Magnetic Resonance(C0024485)	8.25
neuropathies(C0442874)	paralyse(C0522224)	8.25
Degenerative (C0029408) polyarthritis,NOS	Bony sclerosis(C0221434)	7.50
Ketoacidosis, NOS(C0220982)	Lupus erythematosus NOS(C0409974)	7.50
Stenosis, Mitral Valve(C0026269)	Ulcer, Peptic(C0030920)	7.50
depression (C0005587) bipolar	Atrioventricularjunctional(C0232208) rhythm	7.50
Colonoscopy, NOS(C0009378)	Polyp, NOS(C0032584)	7.50

Table A.7: MayoSRS set 3 (SAB : SNOMEDCT, REL : PAR, DIR : U)

Source CUI	Destination CUI	SR
Asthenia NOS(C0004093)	neuropathies(C0442874)	7.50
Injection(C1533685)	Hydrarthrosis, NOS(C1253936)	7.50
(procedure)		
Pneumoniae(C0032285)	Infiltration, NOS(C0332448)	6.75
T wave feature(C0429103)	Infarction, Myocardial(C0027051)	6.75
Sinusitis NOS(C0037199)	Sinusoidal(C0442041)	6.75
Rheumatoid(C0003873)	Arthroscopy, NOS(C0003904)	6.75
arthritis NOS		
Prothrombin(C0033706)	Syringe, NOS(C0039142)	6.75
cortisones(C0010137)	Family history:(C1563292)	6.75
	Osteoporosis	
Dyspareunia (female)(C0013394)	Ovulations(C0029965)	6.75
splinting hand(C0409162)	splinter hemorrhages(C0333286)	6.75
Meniscus structure(C0224498)	Degenerative(C0029408)	6.00
of joint	polyarthritis,NOS	
Ligament rupture(C0262538)	Pointes, Torsades de(C0040479)	6.00
leukaemias(C0023418)	cells stems(C0038250)	6.00
Hallucinations NOS(C0018524)	Disorders, Psychotic(C0033975)	6.00
compression spinal cord(C0037926)	caring wound(C0886052)	6.00
Hepatitides, Autoimmune(C0241910)	Obstetric Labor,(C0022876)	6.00
	Premature	
Antinuclear Antibody(C0003243)	autoimmune reactions(C0443146)	6.00
cirrhosis cryptogenic(C0267809)	gastrins(C0376180)	5.25
Uveitis NOS(C0042164)	Antigen, HLA-B27(C0019740)	5.25
Antibodies, Antiphospholipid(C0162595)	Acne NOS(C0702166)	4.50
Malignant Prostate(C0376358)	Acid Phosphatase(C0001109)	4.50
Neoplasm		
Confusion(C0009676)	Delusion, NOS(C0011253)	4.50
Scleroderma NOS(C0011644)	Scleroderma, Systemic(C0036421)	-1.00
Nodule, Rheumatoid(C0035450)	PULMONARY NODULE(C0034079)	-1.00
Stenosis,(C0003507)	Calcification, Physiologic(C0006660)	-1.00
Aortic Valve		
Rheumatoid arthritis NOS(C0003873)	(C1396859)	-1.00
Walking difficulties(C0311394)	Antalgic gait(C0231685)	-1.00
Disorders, Deglutition(C0011168)	hypomotility(C0679317)	-1.00
Arrhythmia, Cardiac(C0003811)	Valve, Mitral(C0026264)	-1.00
(C2267026)	hyperlipidaemias(C0020473)	-1.00
Aneurysm, Cerebral(C0917996)	Pulmonary Embolisms(C0034065)	-1.00

Table A.8: MayoSRS Big subset 1 test set with key

Source CUI	Destination CUI	SR
Rheumatoid arthritis (C0003873)	Nodule, Rheumatoid (C0035450)	7.08
Ileitis, NOS (C0020877)	Crohn's disease NOS (C0010346)	6.85
Stomatitis NOS (C0038362)	Ulcer, Oral (C0149745)	6.85
Walking difficulties (C0311394)	Antalgic gait (C0231685)	6.69
Hallucinations NOS (C0018524)	Disorders, Psychotic (C0033975)	6.00
joint stiffness (C0457086)	Rheumatoid arthritis (C0003873)	5.69
Diarrhea NOS (C0011991)	COLITIS (NOS) (C0009319)	5.54
Colonoscopy, NOS (C0009378)	Polyp, NOS (C0032584)	5.46
Pain, Back (C0004604)	Stenosis, Spinal (C0037944)	5.31
Paraneoplastic (C0030472)	Malignantneoplasm (C1306459)	5.31
Urticaria NOS (C0042109)	Butterfly rash (C0277942)	5.31
T wave feature (C0429103)	Infarction, Myocardial (C0027051)	5.23
Dyspepsia, NOS (C0013395)	Ulcer, Peptic (C0030920)	5.23
Cerebrovascular accident, NOS (C0038454)	Hemipareses (C0018989)	5.00
phenomenon raynauds (C0034735)	Ischaemia, NOS (C0022116)	4.85
Pneumoniae (C0032285)	Infiltration, NOS (C0332448)	4.85
Calculi, Kidney (C0022650)	Ureteral Obstructions (C0041956)	4.69
Temporal Arteritides (C1956391)	Headache, NOS (C0018681)	4.69
Myopathies (C0026848)	Dermatomyositides (C0011633)	4.46
Deglutition, NOS (C0011167)	Peristalsis, NOS (C0031133)	4.38
Walking (activity) (C0080331)	climbing stair (C0432601)	4.38
Gastrostomy, NOS (C0017196)	Malnutrition NOS (C0162429)	4.31
Brain Stems (C0006121)	Entire cranial nerve (C1269897)	4.23
Asthenia NOS (C0004093)	neuropathies (C0442874)	4.08
Dyspnoea, NOS (C0013404)	Tachypnoea (C0231835)	4.08
Cavitation, NOS (C1510420)	Tuberculosis NOS (C0041296)	4.08
Osteophytes (C1956089)	heberdens nodes (C0018862)	4.00
Peripheral oedema (C0085649)	Oedema - pulmonary (C0034063)	3.92
Ulcer, Leg (C0023223)	Varicosities (C0042345)	3.92
drawers sign (C0231736)	pain in knee (C0231749)	3.77
Disorders, Deglutition (C0011168)	hypomotility (C0679317)	3.77
Rheumatoid arthritis NOS (C0003873)	Arthroscopy, NOS (C0003904)	3.77
Injection (procedure) (C1533685)	Hydrarthrosis, NOS (C1253936)	3.31
Seizure NOS (C0036572)	Headache, NOS (C0018681)	3.31
Pulmonary Embolisms (C0034065)	Pneumoniae (C0032285)	3.23
Malaria NOS (C0024530)	Amoebiasis NOS (C0002438)	3.15
Vasculitis, NOS (C0042384)	Thrombosis, NOS (C0040053)	3.00
Penicillin (C0030842)	Hypersensitivity NOS (C0020517)	3.00
Degenerative polyarthritis (C0029408)	Bony sclerosis (C0221434)	2.85
cortisones (C0010137)	history: Osteoporosis (C1563292)	2.85
neuropathies (C0442874)	paralyse (C0522224)	2.85
Sinusitis NOS (C0037199)	Sinusoidal (C0442041)	2.85

Table A.9: MayoSRS test set Big subset 2 with key

Source CUI	Destination CUI	SR
Pulmonary Embolisms (C0034065)	Hemoptysis NOS (C0019079)	2.77
Meniscus structure joint (C0224498)	polyarthritis (C0029408)	2.62
oedemas (C0013604)	Rate, Glomerular Filtration (C0017654)	2.38
Dyspareunia (female) (C0013394)	Ovulations (C0029965)	2.31
Uveitis NOS (C0042164)	Antigen, HLA-B27 (C0019740)	2.23
immunisations (C0020971)	Immunologic Deficiency (C0021051)	2.15
Laxity, NOS (C0332536)	Syndrome, Marfan (C0024796)	2.08
corneal ulcers (C0010043)	Ulcer, Pressure (C0011127)	2.00
Dysgeusias (C0013378)	Deglutition, NOS (C0011167)	1.92
Prothrombin (C0033706)	Syringe, NOS (C0039142)	1.77
congestive failures (C0018802)	Portal Hypertensions (C0020541)	1.69
Systemic infections (C0243026)	Hypotension NOS (C0020649)	1.69
Malignant Prostate N (C0376358)	Acid Phosphatase (C0001109)	1.69
HAEMATEMESIS (C0018926)	Xerostomia (C0043352)	1.54
Erythema NOS (C0041834)	Osteoporosis NOS (C0029456)	1.54
Sarcoidosis, NOS (C0036202)	Vitamin D, NOS (C0042866)	1.46
Perseveration (C0233651)	Ulcers, Venous Stasis (C1527356)	1.31
Ketoacidosis (C0220982)	Lupus erythematosus (C0409974)	1.23
Myeloma-Multiple (C0026764)	Depressive disorder NOS (C0011581)	1.08
Antibodies, Antiphospholipid (C0162595)	Acne NOS (C0702166)	1.08
Chronic Obstructive Airways (C0024117)	Halitoses (C0018520)	1.08
Ligament rupture (C0262538)	Pointes, Torsades de (C0040479)	1.08
splinting hand (C0409162)	splinter hemorrhages (C0333286)	1.00
Diabetes mellitus NOS (C0011849)	Polyp, NOS (C0032584)	1.00
Portal Hypertensions (C0020541)	melanocytic naevi (C0027962)	1.00
Agents, Contraceptive (C0009871)	Contraindicated (C1444657)	1.00
Lymphoid hyperplasia, NOS (C0333997)	Xerostomia (C0043352)	1.00
Sodium (NOS) (C0037473)	Imaging, Magnetic Resonance (C0024485)	1.00
Diabetes mellitus NOS (C0011849)	Nodule, Rheumatoid (C0035450)	1.00
Stenosis, Tracheal (C0040583)	Malignant Tumor (C0007102)	1.00
cirrhosis cryptogenic (C0267809)	gastrins (C0376180)	1.00
heart failures (C0018801)	Disorders, Psychotic (C0033975)	1.00
rectal polyps (C0034887)	Neoplasms, Laryngeal (C0023055)	1.00
depression bipolar (C0005587)	Atrioventricularjunctional (C0232208)	1.00
Stenosis, Mitral Valve (C0026269)	Ulcer, Peptic (C0030920)	1.00
Adult respiratory syndrome (C0035222)	Cellulitis (C0007642)	1.00
compression spinal cord (C0037926)	caring wound (C0886052)	1.00
Spinal stenosis of cervical (C0158280)	liver cirrhosis (C0023891)	1.00
Hepatitides, Autoimmune (C0241910)	Obstetric Labor (C0022876)	1.00
Small Cell Carcinoma (C0149925)	DiabetesMellitus (C0011860)	1.00
Dyspnoea, NOS (C0013404)	Myalgia NOS (C0231528)	1.00
Fibrosis, Pulmonary (C0034069)	Appendicitis NOS (C0003615)	1.00

Table A.10: MiniMayoSRS big subset (SAB:SNOMEDCT, REL:PAR, DIR:U)

Source CUI	Destination CUI	SR
Renal failure (C0035078)	Kidney failure (C0035078)	20.00
Abortion (C0156543)	Miscarriage (C0000786)	19.00
Heart (C0018787)	Myocardium (C0027061)	17.00
Pulmonary brosis (C0034069)	Lung cancer (C0242379)	12.75
Brain tumor (C0006118)	Intracranial hemorrhage (C0151699)	12.75
Antibiotic (C0003232)	Allergy (C0020517)	11.25
Pulmonary embolus (C0034065)	Myocardial infarction (C0027051)	11.25
Depression (C0011581)	Cellulitis (C0007642)	11.25
Multiple sclerosis (C0026769)	Psychosis (C0033975)	11.25
Congestive heart failure (C0018802)	Pulmonary edema (C0034063)	10.50
Diarrhea (C0011991)	Stomach cramps (C0344375)	10.50
Mitral stenosis (C0026269)	Atrial brillation (C0004238)	10.50
Rheumatoid arthritis (C0003873)	Lupus (C0409974)	10.50
Carpal tunnel syndrome (C0007286)	Osteoarthritis (C0029408)	9.75
Lymphoid hyperplasia (C0333997)	Laryngeal cancer (C0007107)	9.75
Diabetes mellitus (C0011849)	Hypertension (C0020538)	9.75
Xerostomia (C0043352)	Alcoholic cirrhosis (C0023891)	9.00
Appendicitis (C0003615)	Osteoporosis (C0029456)	9.00
Peptic ulcer disease (C0030920)	Myopia (C0027092)	8.25
Cortisone (C0010137)	Total knee replacement (C0086511)	7.50
Acne (C0702166)	Syringe (C0039142)	6.00
Stroke (C0038454)	Infarct (C0021308)	5.25
Cholangiocarcinoma (C0206698)	Colonoscopy (C0009378)	4.50

Table A.11: MayoSRS Big subset 1 (SAB : SNOMEDCT, REL : PAR, DIR : U)

Source CUI	Destination CUI	SR
Myopathies (C0026848)	Dermatomyositides (C0011633)	17.00
Peripheral oedema (C0085649)	Oedema - pulmonary NOS (C0034063)	12.75
drawers sign (C0231736)	pain in knee (C0231749)	12.75
Deglutition, NOS (C0011167)	Peristalsis, NOS (C0031133)	12.75
Vasculitis, NOS (C0042384)	Thrombosis, NOS (C0040053)	12.75
Malaria NOS (C0024530)	Amoebiasis NOS (C0002438)	12.75
corneal ulcers (C0010043)	Ulcer, Pressure (C0011127)	12.75
Calculi, Kidney (C0022650)	Ureteral Obstructions (C0041956)	12.00
Pulmonary Embolisms (C0034065)	Hemoptysis NOS (C0019079)	12.00
Temporal Arteritides (C1956391)	Headache, NOS (C0018681)	12.00
Ileitis, NOS (C0020877)	Crohn's disease NOS (C0010346)	12.00
Stomatitis NOS (C0038362)	Ulcer, Oral (C0149745)	12.00
Ulcer, Leg (C0023223)	Varicosities (C0042345)	12.00
Diabetes mellitus NOS (C0011849)	Polyp, NOS (C0032584)	11.25
congestive failures (C0018802)	Portal Hypertensions (C0020541)	11.25
Pain, Back (C0004604)	Stenosis, Spinal (C0037944)	11.25
Rheumatoid arthritis NOS (C0003873)	Nodule, Rheumatoid (C0035450)	11.25
Dyspnoea, NOS (C0013404)	Tachypnoea (C0231835)	11.25
Syndromes, Paraneoplastic (C0030472)	Malignantneoplasm (C1306459)	11.25
Pulmonary Embolisms (C0034065)	Pneumoniae (C0032285)	11.25
Urticaria NOS (C0042109)	Butterfly rash (C0277942)	11.25
Seizure NOS (C0036572)	Headache, NOS (C0018681)	11.25
Adult respiratory syndrome (C0035222)	Cellulitis, NOS (C0007642)	11.25
Agents, Contraceptive (C0009871)	Contraindicated (C1444657)	10.50
joint stiffness (C0457086)	Rheumatoid arthritis (C0003873)	10.50
Dyspepsia, NOS (C0013395)	Ulcer, Peptic (C0030920)	10.50
Diabetes mellitus NOS (C0011849)	Nodule, Rheumatoid (C0035450)	10.50
Dyspnoea, NOS (C0013404)	Myalgia NOS (C0231528)	10.50
Erythema NOS (C0041834)	Osteoporosis NOS (C0029456)	10.50
Portal Hypertensions (C0020541)	melanocytic naevi (C0027962)	9.75
Cerebrovascular accident, NOS (C0038454)	Hemipareses (C0018989)	9.75
Brain Stems (C0006121)	Entire cranial nerve (C1269897)	9.75
Diarrhea NOS (C0011991)	COLITIS (NOS) (C0009319)	9.75
Systemic infections (C0243026)	Hypotension NOS (C0020649)	9.75
oedemas (C0013604)	Rate, Glomerular Filtration (C0017654)	9.75
rectal polyps (C0034887)	Neoplasms, Laryngeal (C0023055)	9.75
Perseveration (C0233651)	Ulcers, Venous Stasis (C1527356)	9.75
Spinal stenosis (C0158280)	liver cirrhosis (C0023891)	9.75
Fibrosis, Pulmonary (C0034069)	Appendicitis NOS (C0003615)	9.75
Osteophytes (C1956089)	heberdens nodes (C0018862)	9.00
immunisations (C0020971)	Immunologic Deficiency (C0021051)	9.00
Laxity, NOS (C0332536)	Syndrome, Marfan (C0024796)	9.00

Table A.12: MayoSRS Big subset 2 (SAB : SNOMEDCT, REL : PAR, DIR : U)

Source CUI	Destination CUI	SR
Lymphoid hyperplasia (C0333997)	Xerostomia (C0043352)	9.00
Sarcoidosis, NOS (C0036202)	Vitamin D, NOS (C0042866)	9.00
Dysgeusias (C0013378)	Deglutition, NOS (C0011167)	9.00
Penicillin (C0030842)	Hypersensitivity NOS (C0020517)	9.00
heart failures (C0018801)	Disorders, Psychotic (C0033975)	9.00
HAEMATEMESIS (C0018926)	Xerostomia (C0043352)	9.00
Chronic Obstructive Disease (C0024117)	Halitoses (C0018520)	9.00
splinting hand (C0409162)	splinter hemorrhages (C0333286)	8.25
Cavitation, NOS (C1510420)	Tuberculosis NOS (C0041296)	8.25
phenomenon raynauds (C0034735)	Ischaemia, NOS (C0022116)	8.25
Gastrostomy, NOS (C0017196)	Malnutrition NOS (C0162429)	8.25
neuropathies (C0442874)	paralyse (C0522224)	8.25
Walking (activity) (C0080331)	climbing stair (C0432601)	8.25
Sodium (NOS) (C0037473)	Imaging, Magnetic Resonance (C0024485)	8.25
Myeloma-Multiple (C0026764)	Depressive disorder NOS (C0011581)	8.25
Stenosis, Tracheal (C0040583)	Malignant Tumor of the Colon (C0007102)	8.25
Small Cell Carcinoma (C0149925)	DiabetesMellitus (C0011860)	8.25
Asthenia NOS (C0004093)	neuropathies (C0442874)	7.50
Degenerative polyarthritis (C0029408)	Bony sclerosis (C0221434)	7.50
Injection (procedure) (C1533685)	Hydrarthrosis, NOS (C1253936)	7.50
Colonoscopy, NOS (C0009378)	Polyp, NOS (C0032584)	7.50
Ketoacidosis, NOS (C0220982)	Lupus erythematosus NOS (C0409974)	7.50
depression bipolar (C0005587)	Atrioventricularjunctional (C0232208)	7.50
Stenosis, Mitral Valve (C0026269)	Ulcer, Peptic (C0030920)	7.50
compression spinal cord (C0037926)	caring wound (C0886052)	7.50
T wave feature (C0429103)	Infarction, Myocardial (C0027051)	6.75
Pneumoniae (C0032285)	Infiltration, NOS (C0332448)	6.75
Rheumatoid arthritis NOS (C0003873)	Arthroscopy, NOS (C0003904)	6.75
cortisones (C0010137)	Family history: Osteoporosis (C1563292)	6.75
Dyspareunia (female) (C0013394)	Ovulations (C0029965)	6.75
Prothrombin (C0033706)	Syringe, NOS (C0039142)	6.75
Sinusitis NOS (C0037199)	Sinusoidal (C0442041)	6.75
Ligament rupture (C0262538)	Pointes, Torsades de (C0040479)	6.75
Hallucinations NOS (C0018524)	Disorders, Psychotic (C0033975)	6.00
Meniscus structure (C0224498)	Degenerative polyarthritis (C0029408)	6.00
Hepatitides, Autoimmune (C0241910)	Obstetric Labor, Premature (C0022876)	6.00
Uveitis NOS (C0042164)	Antigen, HLA-B27 (C0019740)	5.25
cirrhosis cryptogenic (C0267809)	gastrins (C0376180)	5.25
Malignant Prostate (C0376358)	Acid Phosphatase (C0001109)	4.50
Antibodies, Antiphospholipid (C0162595)	Acne NOS (C0702166)	4.50
Walking difficulties (C0311394)	Antalgic gait (C0231685)	-1.00
Disorders, Deglutition (C0011168)	hypomotility (C0679317)	-1.00

Table A.13: MiniMayoSRS big subset (SAB:SNOMEDCT, REL : PAR,RB,RN,RO, DIR:U,H,H,H) Cost of H link is 3

Source CUI	Destination CUI	SR
Renal failure (C0035078)	Kidney failure (C0035078)	20.00
Abortion (C0156543)	Miscarriage (C0000786)	19.00
Heart (C0018787)	Myocardium (C0027061)	17.00
Pulmonary brosis (C0034069)	Lung cancer (C0242379)	12.75
Brain tumor (C0006118)	Intracranial hemorrhage (C0151699)	12.75
Stroke (C0038454)	Infarct (C0021308)	12.00
Antibiotic (C0003232)	Allergy (C0020517)	11.25
Pulmonary embolus (C0034065)	Myocardial infarction (C0027051)	11.25
Depression (C0011581)	Cellulitis (C0007642)	11.25
Multiple sclerosis (C0026769)	Psychosis (C0033975)	11.25
Congestive heart failure (C0018802)	Pulmonary edema (C0034063)	10.50
Diarrhea (C0011991)	Stomach cramps (C0344375)	10.50
Mitral stenosis (C0026269)	Atrial brillation (C0004238)	10.50
Rheumatoid arthritis (C0003873)	Lupus (C0409974)	10.50
Carpal tunnel syndrome (C0007286)	Osteoarthritis (C0029408)	9.75
Lymphoid hyperplasia (C0333997)	Laryngeal cancer (C0007107)	9.75
Diabetes mellitus (C0011849)	Hypertension (C0020538)	9.75
Xerostomia (C0043352)	Alcoholic cirrhosis (C0023891)	9.00
Appendicitis (C0003615)	Osteoporosis (C0029456)	9.00
Peptic ulcer disease (C0030920)	Myopia (C0027092)	8.25
Cortisone (C0010137)	Total knee replacement (C0086511)	7.50
Cholangiocarcinoma (C0206698)	Colonoscopy (C0009378)	7.50
Acne (C0702166)	Syringe (C0039142)	6.00

Table A.14: MayoSRS Big subset 1 (SAB : SNOMEDCT, REL : PAR,RB,RN,RO, DIR : U,H,H,H), Cost of H link is 3

Source CUI	Destination CUI	SR
Myopathies (C0026848)	Dermatomyositides (C0011633)	17.00
Rheumatoid arthritis NOS (C0003873)	Arthroscopy, NOS (C0003904)	14.00
Peripheral oedema (C0085649)	Oedema - pulmonary NOS (C0034063)	12.75
drawers sign (C0231736)	pain in knee (C0231749)	12.75
Deglutition, NOS (C0011167)	Peristalsis, NOS (C0031133)	12.75
Vasculitis, NOS (C0042384)	Thrombosis, NOS (C0040053)	12.75
Malaria NOS (C0024530)	Amoebiasis NOS (C0002438)	12.75
corneal ulcers (C0010043)	Ulcer, Pressure (C0011127)	12.75
Calculi, Kidney (C0022650)	Ureteral Obstructions (C0041956)	12.00
Osteophytes (C1956089)	heberdens nodes (C0018862)	12.00
Pulmonary Embolisms (C0034065)	Hemoptysis NOS (C0019079)	12.00
Temporal Arteritides (C1956391)	Headache, NOS (C0018681)	12.00
Ileitis, NOS (C0020877)	Crohn's disease NOS (C0010346)	12.00
Stomatitis NOS (C0038362)	Ulcer, Oral (C0149745)	12.00
Ulcer, Leg (C0023223)	Varicosities (C0042345)	12.00
Diabetes mellitus NOS (C0011849)	Polyp, NOS (C0032584)	11.25
congestive failures heart (C0018802)	Portal Hypertensions (C0020541)	11.25
Pain, Back (C0004604)	Stenosis, Spinal (C0037944)	11.25
Rheumatoid arthritis NOS (C0003873)	Rheumatoid (C0035450)	11.25
Dyspnoea, NOS (C0013404)	Tachypnoea (C0231835)	11.25
Syndromes, Paraneoplastic (C0030472)	Malignantneoplasm, primary (C1306459)	11.25
Pulmonary Embolisms (C0034065)	Pneumoniae (C0032285)	11.25
Urticaria NOS (C0042109)	Butterfly rash (C0277942)	11.25
Seizure NOS (C0036572)	Headache, NOS (C0018681)	11.25
Adult respiratory distress syndrome (C0035222)	Cellulitis, NOS (C0007642)	11.25
Pneumoniae (C0032285)	Infiltration, NOS (C0332448)	11.00
Colonoscopy, NOS (C0009378)	Polyp, NOS (C0032584)	11.00
Agents, Contraceptive (C0009871)	Contraindicated (C1444657)	10.50
joint morning stiffness (C0457086)	Rheumatoid arthritis NOS (C0003873)	10.50
Brain Stems (C0006121)	Entire cranial nerve (C1269897)	10.50
Dyspepsia, NOS (C0013395)	Ulcer, Peptic (C0030920)	10.50
Meniscus structure of joint (C0224498)	Degenerative polyarthritis,NOS (C0029408)	10.50
Diabetes mellitus NOS (C0011849)	Nodule, Rheumatoid (C0035450)	10.50
Dyspnoea, NOS (C0013404)	Myalgia NOS (C0231528)	10.50
Erythema NOS (C0041834)	Osteoporosis NOS (C0029456)	10.50
Portal Hypertensions (C0020541)	melanocytic naevi (C0027962)	9.75
Cerebrovascular accident, NOS (C0038454)	Hemipareses (C0018989)	9.75
Diarrhea NOS (C0011991)	COLITIS (NOS) (C0009319)	9.75
Systemic infections (C0243026)	Hypotension NOS (C0020649)	9.75
oedemas (C0013604)	Rate, Glomerular Filtration (C0017654)	9.75
rectal polyps (C0034887)	Neoplasms, Laryngeal (C0023055)	9.75
Perseveration (C0233651)	Ulcers, Venous Stasis (C1527356)	9.75

Table A.15: MayoSRS Big subset 2 (SAB : SNOMEDCT, REL : PAR,RB,RN,RO, DIR : U,H,H,H), Cost of H link is 3

Source CUI	Destination CUI	SR
Spinal stenosis (C0158280)	liver cirrhosis (C0023891)	9.75
Fibrosis, Pulmonary (C0034069)	Appendicitis NOS (C0003615)	9.75
immunisations (C0020971)	Immunologic Deficiency (C0021051)	9.00
Laxity, NOS (C0332536)	Syndrome, Marfan (C0024796)	9.00
Lymphoid hyperplasia (C0333997)	Xerostomia (C0043352)	9.00
Sarcoidosis, NOS (C0036202)	Vitamin D, NOS (C0042866)	9.00
Dysgeusias (C0013378)	Deglutition, NOS (C0011167)	9.00
Penicillin (C0030842)	Hypersensitivity NOS (C0020517)	9.00
heart failures (C0018801)	Disorders, Psychotic (C0033975)	9.00
HAEMATEMESIS (C0018926)	Xerostomia (C0043352)	9.00
Chronic Obstructive Disease (C0024117)	Halitoses (C0018520)	9.00
splinting hand (C0409162)	splinter hemorrhages (C0333286)	8.25
Cavitation, NOS (C1510420)	Tuberculosis NOS (C0041296)	8.25
phenomenon raynauds (C0034735)	Ischaemia, NOS (C0022116)	8.25
Gastrostomy, NOS (C0017196)	Malnutrition NOS (C0162429)	8.25
neuropathies (C0442874)	paralyse (C0522224)	8.25
Walking (activity) (C0080331)	climbing stair (C0432601)	8.25
Sodium (NOS) (C0037473)	Imaging, Magnetic Resonance (C0024485)	8.25
Myeloma-Multiple (C0026764)	Depressive disorder (C0011581)	8.25
Stenosis, Tracheal (C0040583)	Malignant Tumor (C0007102)	8.25
Small Cell Carcinoma (C0149925)	DiabetesMellitus (C0011860)	8.25
Asthenia NOS (C0004093)	neuropathies (C0442874)	7.50
Degenerative polyarthritis,NOS (C0029408)	Bony sclerosis (C0221434)	7.50
Injection (procedure) (C1533685)	Hydrarthrosis, NOS (C1253936)	7.50
Ketoacidosis, NOS (C0220982)	Lupus erythematosus (C0409974)	7.50
depression bipolar (C0005587)	Atrioventricularjunctional (C0232208)	7.50
Stenosis, Mitral Valve (C0026269)	Ulcer, Peptic (C0030920)	7.50
compression spinal cord (C0037926)	caring wound (C0886052)	7.50
T wave feature (C0429103)	Infarction, Myocardial (C0027051)	6.75
cortisones (C0010137)	Family history: Osteoporosis (C1563292)	6.75
Dyspareunia (female) (C0013394)	Ovulations (C0029965)	6.75
Prothrombin (C0033706)	Syringe, NOS (C0039142)	6.75
Sinusitis NOS (C0037199)	Sinusoidal (C0442041)	6.75
Ligament rupture (C0262538)	Pointes, Torsades de (C0040479)	6.75
Hallucinations NOS (C0018524)	Disorders, Psychotic (C0033975)	6.00
Hepatitides, Autoimmune (C0241910)	Obstetric Labor (C0022876)	6.00
Uveitis NOS (C0042164)	Antigen, HLA-B27 (C0019740)	5.25
cirrhosis cryptogenic (C0267809)	gastrins (C0376180)	5.25
Walking difficulties (C0311394)	Antalgic gait (C0231685)	5.00
Malignant Prostate (C0376358)	Acid Phosphatase (C0001109)	4.50
Antibodies, Antiphospholipid (C0162595)	Acne NOS (C0702166)	4.50
Disorders, Deglutition (C0011168)	hypomotility (C0679317)	-1.00

Table A.16: MiniMayoSRS big subset (SAB:SNOMEDCT, REL : PAR,RB,RN,RO, DIR:U,H,H,H) Cost of H link is 2

Source CUI	Destination CUI	SR
Renal failure (C0035078)	Kidney failure (C0035078)	20.00
Abortion (C0156543)	Miscarriage (C0000786)	19.00
Heart (C0018787)	Myocardium (C0027061)	17.00
Stroke (C0038454)	Infarct (C0021308)	12.75
Pulmonary brosis (C0034069)	Lung cancer (C0242379)	12.75
Brain tumor (C0006118)	Intracranial hemorrhage (C0151699)	12.75
Antibiotic (C0003232)	Allergy (C0020517)	11.25
Pulmonary embolus (C0034065)	Myocardial infarction (C0027051)	11.25
Depression (C0011581)	Cellulitis (C0007642)	11.25
Multiple sclerosis (C0026769)	Psychosis (C0033975)	11.25
Congestive heart failure (C0018802)	Pulmonary edema (C0034063)	10.50
Diarrhea (C0011991)	Stomach cramps (C0344375)	10.50
Mitral stenosis (C0026269)	Atrial brillation (C0004238)	10.50
Rheumatoid arthritis (C0003873)	Lupus (C0409974)	10.50
Carpal tunnel syndrome (C0007286)	Osteoarthritis (C0029408)	9.75
Lymphoid hyperplasia (C0333997)	Laryngeal cancer (C0007107)	9.75
Cholangiocarcinoma (C0206698)	Colonoscopy (C0009378)	9.75
Diabetes mellitus (C0011849)	Hypertension (C0020538)	9.75
Xerostomia (C0043352)	Alcoholic cirrhosis (C0023891)	9.00
Appendicitis (C0003615)	Osteoporosis (C0029456)	9.00
Peptic ulcer disease (C0030920)	Myopia (C0027092)	8.25
Cortisone (C0010137)	Total knee replacement (C0086511)	7.50
Acne (C0702166)	Syringe (C0039142)	6.75

Table A.17: MayoSRS Big subset 1 (SAB : SNOMEDCT, REL : PAR,RB,RN,RO, DIR : U,H,H,H), Cost of H link is 2

Source CUI	Destination CUI	SR
Myopathies (C0026848)	Dermatomyositides (C0011633)	17.00
Rheumatoid arthritis NOS (C0003873)	Arthroscopy, NOS (C0003904)	16.00
Pneumoniae (C0032285)	Infiltration, NOS (C0332448)	14.00
Diarrhea NOS (C0011991)	COLITIS (NOS) (C0009319)	14.00
Colonoscopy, NOS (C0009378)	Polyp, NOS (C0032584)	14.00
Peripheral oedema (C0085649)	Oedema - pulmonary NOS (C0034063)	12.75
Osteophytes (C1956089)	heberdens nodes (C0018862)	12.75
drawers sign (C0231736)	pain in knee (C0231749)	12.75
Deglutition, NOS (C0011167)	Peristalsis, NOS (C0031133)	12.75
Vasculitis, NOS (C0042384)	Thrombosis, NOS (C0040053)	12.75
Malaria NOS (C0024530)	Amoebiasis NOS (C0002438)	12.75
Ulcer, Leg (C0023223)	Varicosities (C0042345)	12.75
corneal ulcers (C0010043)	Ulcer, Pressure (C0011127)	12.75
Calculi, Kidney (C0022650)	Ureteral Obstructions (C0041956)	12.00
Pulmonary Embolisms (C0034065)	Hemoptysis NOS (C0019079)	12.00
Temporal Arteritides (C1956391)	Headache, NOS (C0018681)	12.00
Ileitis, NOS (C0020877)	Crohn's disease NOS (C0010346)	12.00
Stomatitis NOS (C0038362)	Ulcer, Oral (C0149745)	12.00
Diabetes mellitus NOS (C0011849)	Polyp, NOS (C0032584)	11.25
congestive failures heart (C0018802)	Portal Hypertensions (C0020541)	11.25
Pain, Back (C0004604)	Stenosis, Spinal (C0037944)	11.25
Rheumatoid arthritis NOS (C0003873)	Nodule, Rheumatoid (C0035450)	11.25
Dyspnoea, NOS (C0013404)	Tachypnoea (C0231835)	11.25
Syndromes, Paraneoplastic (C0030472)	Malignantneoplasm, primary (C1306459)	11.25
Pulmonary Embolisms (C0034065)	Pneumoniae (C0032285)	11.25
Urticaria NOS (C0042109)	Butterfly rash (C0277942)	11.25
Seizure NOS (C0036572)	Headache, NOS (C0018681)	11.25
Meniscus structure of joint (C0224498)	Degenerative polyarthritis,NOS (C0029408)	11.25
Adult respiratory distress syndrome (C0035222)	Cellulitis, NOS (C0007642)	11.25
Fibrosis, Pulmonary (C0034069)	Appendicitis NOS (C0003615)	11.25
Agents, Contraceptive (C0009871)	Contraindicated (C1444657)	10.50
joint morning stiffness (C0457086)	Rheumatoid arthritis NOS (C0003873)	10.50
Dyspepsia, NOS (C0013395)	Ulcer, Peptic (C0030920)	10.50
Diabetes mellitus NOS (C0011849)	Nodule, Rheumatoid (C0035450)	10.50
Dyspnoea, NOS (C0013404)	Myalgia NOS (C0231528)	10.50
Erythema NOS (C0041834)	Osteoporosis NOS (C0029456)	10.50
Brain Stems (C0006121)	Entire cranial nerve (C1269897)	10.50
Walking difficulties (C0311394)	Antalgic gait (C0231685)	10.00
cirrhosis cryptogenic (C0267809)	gastrins (C0376180)	10.00
Ligament rupture (C0262538)	Pointes, Torsades de (C0040479)	10.00
Portal Hypertensions (C0020541)	melanocytic naevi (C0027962)	9.75
Degenerative polyarthritis,NOS (C0029408)	Bony sclerosis (C0221434)	9.75

Table A.18: MayoSRS Big subset 2 (SAB : SNOMEDCT, REL : PAR,RB,RN,RO, DIR : U,H,H,H), Cost of H link is 2

Source CUI	Destination CUI	SR
Cerebrovascular accident, NOS (C0038454)	Hemipareses (C0018989)	9.75
Systemic infections (C0243026)	Hypotension NOS (C0020649)	9.75
oedemas (C0013604)	Rate, Glomerular Filtration (C0017654)	9.75
Penicillin (C0030842)	Hypersensitivity NOS (C0020517)	9.75
rectal polyps (C0034887)	Neoplasms, Laryngeal (C0023055)	9.75
Perseveration (C0233651)	Ulcers, Venous Stasis (C1527356)	9.75
Spinal stenosis (C0158280)	liver cirrhosis (C0023891)	9.75
immunisations (C0020971)	Immunologic Deficiency (C0021051)	9.00
Laxity, NOS (C0332536)	Syndrome, Marfan (C0024796)	9.00
Lymphoid hyperplasia, NOS (C0333997)	Xerostomia (C0043352)	9.00
Sarcoidosis, NOS (C0036202)	Vitamin D, NOS (C0042866)	9.00
Dysgeusias (C0013378)	Deglutition, NOS (C0011167)	9.00
heart failures (C0018801)	Disorders, Psychotic (C0033975)	9.00
HAEMATEMESIS (C0018926)	Xerostomia (C0043352)	9.00
Chronic Obstructive Disease (C0024117)	Halitoses (C0018520)	9.00
splinting hand (C0409162)	splinter hemorrhages (C0333286)	8.25
T wave feature (C0429103)	Infarction, Myocardial (C0027051)	8.25
Cavitation, NOS (C1510420)	Tuberculosis NOS (C0041296)	8.25
phenomenon raynauds (C0034735)	Ischaemia, NOS (C0022116)	8.25
Gastrostomy, NOS (C0017196)	Malnutrition NOS (C0162429)	8.25
neuropathies (C0442874)	paralyse (C0522224)	8.25
Walking (activity) (C0080331)	climbing stair (C0432601)	8.25
Sodium (NOS) (C0037473)	Imaging, Magnetic Resonance (C0024485)	8.25
Myeloma-Multiple (C0026764)	Depressive disorder NOS (C0011581)	8.25
Stenosis, Tracheal (C0040583)	Malignant Tumor of the Colon (C0007102)	8.25
Small Cell Carcinoma (C0149925)	DiabetesMellitus (C0011860)	8.25
Asthenia NOS (C0004093)	neuropathies (C0442874)	7.50
Injection (procedure) (C1533685)	Hydrarthrosis, NOS (C1253936)	7.50
Ketoacidosis, NOS (C0220982)	Lupus erythematosus NOS (C0409974)	7.50
depression bipolar (C0005587)	Atrioventricularjunctional (C0232208)	7.50
Stenosis, Mitral Valve (C0026269)	Ulcer, Peptic (C0030920)	7.50
compression spinal cord (C0037926)	caring wound (C0886052)	7.50
Uveitis NOS (C0042164)	Antigen, HLA-B27 (C0019740)	6.75
cortisones (C0010137)	Family history: Osteoporosis (C1563292)	6.75
Dyspareunia (female) (C0013394)	Ovulations (C0029965)	6.75
Prothrombin (C0033706)	Syringe, NOS (C0039142)	6.75
Sinusitis NOS (C0037199)	Sinusoidal (C0442041)	6.75
Hallucinations NOS (C0018524)	Disorders, Psychotic (C0033975)	6.00
Hepatitides, Autoimmune (C0241910)	Obstetric Labor, Premature (C0022876)	6.00
Malignant Prostate Neoplasm (C0376358)	Acid Phosphatase (C0001109)	4.50
Antibodies, Antiphospholipid (C0162595)	Acne NOS (C0702166)	4.50
Disorders, Deglutition (C0011168)	hypomotility (C0679317)	-1.00

Table A.19: MiniMayoSRS big subset (SAB:SNOMEDCT, REL : PAR,RB,RN,RO, DIR:U,H,H,H) Cost of H link is 1

Source CUI	Destination CUI	SR
Renal failure (C0035078)	Kidney failure (C0035078)	20.00
Abortion (C0156543)	Miscarriage (C0000786)	19.00
Heart (C0018787)	Myocardium (C0027061)	18.00
Pulmonary embolus (C0034065)	Myocardial infarction (C0027051)	16.00
Carpal tunnel syndrome (C0007286)	Osteoarthritis (C0029408)	16.00
Stroke (C0038454)	Infarct (C0021308)	13.50
Pulmonary brosis (C0034069)	Lung cancer (C0242379)	12.75
Brain tumor (C0006118)	Intracranial hemorrhage (C0151699)	12.75
Lymphoid hyperplasia (C0333997)	Laryngeal cancer (C0007107)	12.00
Cholangiocarcinoma (C0206698)	Colonoscopy (C0009378)	12.00
Antibiotic (C0003232)	Allergy (C0020517)	11.25
Rheumatoid arthritis (C0003873)	Lupus (C0409974)	11.25
Depression (C0011581)	Cellulitis (C0007642)	11.25
Multiple sclerosis (C0026769)	Psychosis (C0033975)	11.25
Xerostomia (C0043352)	Alcoholic cirrhosis (C0023891)	11.25
Congestive heart failure (C0018802)	Pulmonary edema (C0034063)	10.50
Diarrhea (C0011991)	Stomach cramps (C0344375)	10.50
Mitral stenosis (C0026269)	Atrial brillation (C0004238)	10.50
Acne (C0702166)	Syringe (C0039142)	10.50
Diabetes mellitus (C0011849)	Hypertension (C0020538)	9.75
Appendicitis (C0003615)	Osteoporosis (C0029456)	9.00
Peptic ulcer disease (C0030920)	Myopia (C0027092)	8.25
Cortisone (C0010137)	Total knee replacement (C0086511)	7.50

Table A.20: MayoSRS Big subset 1 (SAB : SNOMEDCT, REL : PAR,RB,RN,RO, DIR : U,H,H,H), Cost of H link is 1

Source CUI	Destination CUI	SR
Rheumatoid arthritis NOS (C0003873)	Arthroscopy, NOS (C0003904)	18.00
Myopathies (C0026848)	Dermatomyositides (C0011633)	17.00
Pneumoniae (C0032285)	Infiltration, NOS (C0332448)	17.00
Diarrhea NOS (C0011991)	COLITIS (NOS) (C0009319)	17.00
Colonoscopy, NOS (C0009378)	Polyp, NOS (C0032584)	17.00
Seizure NOS (C0036572)	Headache, NOS (C0018681)	16.00
rectal polyps (C0034887)	Neoplasms, Laryngeal (C0023055)	16.00
Walking difficulties (C0311394)	Antalgic gait (C0231685)	15.00
Penicillin (C0030842)	Hypersensitivity NOS (C0020517)	15.00
cirrhosis cryptogenic (C0267809)	gastrins (C0376180)	15.00
Ligament rupture (C0262538)	Pointes, Torsades de (C0040479)	15.00
Laxity, NOS (C0332536)	Syndrome, Marfan (C0024796)	14.00
Stenosis, Tracheal (C0040583)	Malignant Tumor of the Colon (C0007102)	14.00
Stenosis, Mitral Valve (C0026269)	Ulcer, Peptic (C0030920)	14.00
Spinal stenosis (C0158280)	liver cirrhosis (C0023891)	14.00
Peripheral oedema (C0085649)	Oedema - pulmonary NOS (C0034063)	13.50
Osteophytes (C1956089)	heberdens nodes (C0018862)	13.50
Ulcer, Leg (C0023223)	Varicosities (C0042345)	13.50
Malignant Prostate Neoplasm (C0376358)	Acid Phosphatase (C0001109)	13.00
Rheumatoid arthritis NOS (C0003873)	Nodule, Rheumatoid (C0035450)	12.75
drawers sign (C0231736)	pain in knee (C0231749)	12.75
Deglutition, NOS (C0011167)	Peristalsis, NOS (C0031133)	12.75
Vasculitis, NOS (C0042384)	Thrombosis, NOS (C0040053)	12.75
Malaria NOS (C0024530)	Amoebiasis NOS (C0002438)	12.75
corneal ulcers (C0010043)	Ulcer, Pressure (C0011127)	12.75
Fibrosis, Pulmonary (C0034069)	Appendicitis NOS (C0003615)	12.75
Brain Stems (C0006121)	Entire cranial nerve (C1269897)	12.75
Calculi, Kidney (C0022650)	Ureteral Obstructions (C0041956)	12.00
Pulmonary Embolisms (C0034065)	Hemoptysis NOS (C0019079)	12.00
Degenerative polyarthritis,NOS (C0029408)	Bony sclerosis (C0221434)	12.00
Temporal Arteritides (C1956391)	Headache, NOS (C0018681)	12.00
Ileitis, NOS (C0020877)	Crohn's disease NOS (C0010346)	12.00
Stomatitis NOS (C0038362)	Ulcer, Oral (C0149745)	12.00
Meniscus structure (C0224498)	Degenerative polyarthritis (C0029408)	12.00
Adult respiratory distress syndrome (C0035222)	Cellulitis, NOS (C0007642)	12.00
Diabetes mellitus NOS (C0011849)	Polyp, NOS (C0032584)	11.25
congestive failures heart (C0018802)	Portal Hypertensions (C0020541)	11.25
Pain, Back (C0004604)	Stenosis, Spinal (C0037944)	11.25
Dyspnoea, NOS (C0013404)	Tachypnoea (C0231835)	11.25
T wave feature (C0429103)	Infarction, Myocardial (C0027051)	11.25
Syndromes, Paraneoplastic (C0030472)	Malignantneoplasm, primary (C1306459)	11.25
Pulmonary Embolisms (C0034065)	Pneumoniae (C0032285)	11.25

Table A.21: MayoSRS Big subset 2 (SAB : SNOMEDCT, REL : PAR,RB,RN,RO, DIR : U,H,H,H), Cost of H link is 1

Source CUI	Destination CUI	SR
Urticaria NOS (C0042109)	Butterfly rash (C0277942)	11.25
Diabetes mellitus NOS (C0011849)	Nodule, Rheumatoid (C0035450)	11.25
Agents, Contraceptive (C0009871)	Contraindicated (C1444657)	10.50
joint morning stiffness (C0457086)	Rheumatoid arthritis (C0003873)	10.50
Cavitation, NOS (C1510420)	Tuberculosis NOS (C0041296)	10.50
Dyspepsia, NOS (C0013395)	Ulcer, Peptic (C0030920)	10.50
Uveitis NOS (C0042164)	Antigen, HLA-B27 (C0019740)	10.50
Sarcoidosis, NOS (C0036202)	Vitamin D, NOS (C0042866)	10.50
Sodium (NOS) (C0037473)	Imaging, Magnetic Resonance (C0024485)	10.50
Dyspnoea, NOS (C0013404)	Myalgia NOS (C0231528)	10.50
Erythema NOS (C0041834)	Osteoporosis NOS (C0029456)	10.50
Portal Hypertensions (C0020541)	melanocytic naevi (C0027962)	9.75
Cerebrovascular accident (C0038454)	Hemipareses (C0018989)	9.75
Systemic infections (C0243026)	Hypotension NOS (C0020649)	9.75
oedemas (C0013604)	Rate, Glomerular Filtration (C0017654)	9.75
neuropathies (C0442874)	paralyse (C0522224)	9.75
Perseveration (C0233651)	Ulcers, Venous Stasis (C1527356)	9.75
immunisations (C0020971)	Immunologic Deficiency (C0021051)	9.00
Lymphoid hyperplasia, NOS (C0333997)	Xerostomia (C0043352)	9.00
Dysgeusias (C0013378)	Deglutition, NOS (C0011167)	9.00
heart failures (C0018801)	Disorders, Psychotic (C0033975)	9.00
HAEMATEMESIS (C0018926)	Xerostomia (C0043352)	9.00
Chronic Obstructive Disease (C0024117)	Halitoses (C0018520)	9.00
splinting hand (C0409162)	splinter hemorrhages (C0333286)	8.25
phenomenon raynauds (C0034735)	Ischaemia, NOS (C0022116)	8.25
Gastrostomy, NOS (C0017196)	Malnutrition NOS (C0162429)	8.25
Walking (activity) (C0080331)	climbing stair (C0432601)	8.25
Myeloma-Multiple (C0026764)	Depressive disorder NOS (C0011581)	8.25
Small Cell Carcinoma (C0149925)	DiabetesMellitus (C0011860)	8.25
Asthenia NOS (C0004093)	neuropathies (C0442874)	7.50
Injection (procedure) (C1533685)	Hydrarthrosis, NOS (C1253936)	7.50
Ketoacidosis, NOS (C0220982)	Lupus erythematosus NOS (C0409974)	7.50
depression bipolar (C0005587)	Atrioventricularjunctional (C0232208)	7.50
compression spinal cord (C0037926)	caring wound (C0886052)	7.50
Dyspareunia (female) (C0013394)	Ovulations (C0029965)	6.75
Prothrombin (C0033706)	Syringe, NOS (C0039142)	6.75
Sinusitis NOS (C0037199)	Sinusoidal (C0442041)	6.75
Hallucinations NOS (C0018524)	Disorders, Psychotic (C0033975)	6.00
Hepatitides, Autoimmune (C0241910)	Obstetric Labor, Premature (C0022876)	6.00
Disorders, Deglutition (C0011168)	hypomotility (C0679317)	-1.00