Unsupervised Learning of Ontology for the Medical Domain

MSc Thesis

Sónia Margarida Ferreira de Bastos

Unsupervised Learning of Ontology for the Medical Domain

DISSERTATION

submitted to University of Beira Interior in partial fulfillment of the requirement for the degree of

MASTER OF SCIENCE

in

INFORMATION SYSTEMS AND TECHNOLOGIES

by

Sónia Margarida Ferreira de Bastos

Department of Computer Science
University of Beira Interior
Covilhã, Portugal
www.di.ubi.pt

HULTIG - Centre of Human Language
Technology and Bioinformatics
Department of Computer Science of the
University of Beira Interior
Covilhã, Portugal
hultig.di.ubi.pt

Cover image: Heraldry of the University of Beira Interior.

Unsupervised Learning of Ontology for the Medical Domain

Author:	Sónia Margarida Ferreira de Bastos
Student number:	m2149
Email:	sonia@hultig.di.ubi.pt
Supervisor:	Prof. Dr. Gaël Dias, UBI, Covilhã

Abstract

Tom Gruber (1993) defines *Ontology* as "*an explicit specification of a conceptualization.*"

Due two the enormous quantity of information available, there is a growing number of applications that perform tasks where lexical-semantic resources are needed, like Information Retrieval, intelligent search or machine translation. This shows that Natural Language Processing is becoming more dependent on semantic information.

One of the main motivations in ontology building is the possibility of knowledge sharing and reuse across different applications. The start point is to fixed a particular domain (like medicine), which is expected to be the base of domain knowledge for a variety of applications. This is a difficult task as the domain knowledge strongly depends on the particular task at hand.

This paper is an approach on ontology learning, for which it was selected the Medical Domain, so that we could have a base to compare and evaluate the resulting ontology.

In our approach, we use different techniques, like Asymmetric Association Measures, clustering algorithm and text rank algorithm, so that we can obtain relations between a set of terms, which are rank by the degree of generality, like the cluster obtained by applying clustering algorithms, with the confidence measure as the values for the similarity matrix, to the set of terms, the generality clusters. Those clusters are then submitted to clustering algorithm, but with Symmetric Conditional Probability values in the similarity matrix, to obtain domain clusters within the generality clusters.

In the future, this ontology may be used in acquisition of Lexical Chains for Text Summarization, as in other Natural Language Processing applications.

Acknowledgments

There are many people, without whose support and contributions this thesis would not have been possible. I am specially grateful to my supervisor, Gaël Dias for his constant support and help, for professional and personal support, and for getting me interested in the subject of Natural Language Processing and for introducing me to the summarization and ontology building fields. I also want to thank to HULTIG research group, in particular Rumen and João Paulo Cordeiro for sharing their work, which had elements necessary for the progress of my work.

I would like to thank all my friends for all their help and support when sometimes I thought I would not make it.

And finally, I want to thank to my mother, Adélia Simão, my boyfriend, Helder, and his family (especially his mother and grandmother) for all their help, support and incentive.

Thank you for all your support!

Sónia Margarida Ferreira de Bastos
Covilhã, Portugal
August, 2009

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Due two the enormous quantity of information available, there is a growing number of applications that perform tasks where lexical-semantic resources are needed, like Information Retrieval, intelligent search or machine translation. This shows that Natural Language Processing is becoming more dependent on semantic information.

Nowadays, researchers are trying to build ontologies, which are models that aim to represent concepts hierarchically, and the relations between them.

The term *ontology* comes from the field of philosophy that is concerned with the study of being or existence.

In the computer science field, Tom Gruber [12] defines *Ontology* as "*an explicit specification of a conceptualization.*".

We can say that an *Ontology* is a formal representation of a set of concepts within a domain and the relationships between them. Ontologies are used in artificial intelligence, the Semantic Web, software engineering, biomedical informatics, library science, and information architecture as a form of knowledge representation about the world or some part of it.

An example of a published *Ontology* is WordNet, in which *Ontology*, in the computer science filed, is defined as "*a rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations*".

Most of the Ontologies developed, or currently in development are not publicly available, which implies research and new implementation of ontologies by researches.

In order to avoid time-consuming human work in its construction and maintenance, we aim to design and, in the future, develop the tools needed to automatically (without supervision) build ontologies without access to existing linguistic resources (thesaurus, lexico-semantic databases, ...), maintaining the process as language-independent as possible.

Although we want to maintain it language-independent, we choose the medical domain

in Enghish as a place to start. Therefore, our work uses UMLS Metathesaurus, so that we can have a base of comparison for our results. This Metathesaurus contains a collection of medical concepts and terms of the various vocabularies and the relationships between them.

It also uses MEDLINE, a bibliographic database containing over 17 million references to journal articles in life sciences and biomedical information, for gathering a set of documents, so we could count the occurrence and the co-occurrence of the terms, which is necessary to calculate the different measures in our work.

This work is divided in four parts. The first describes the resources we used: (1) UMLS and (2) MEDLINE. The next parts describes all the steps we made in building the Ontology - the association measures, the rank and cluster algorithms, the requirements for their implementation and the problems we found.

Afterwards we make a reference to related work in building ontologies, but for Portuguese language.

For last we present some conclusions and future work.

Chapter 2

Resources

2.1 UMLS - Unified Medical Language System

UMLS is a compendium of vocabularies in several languages in the biomedical sciences and the relations between them (this relations are only available for English). It is produced and distributed by the National Library of Medicine (NLM) - National Institutes of Medicine - USA. All the information about UMLS is available at http://www.nlm.nih.gov/research/umls/.

The use of UMLS Knowledge Sources and the associated software is free of charge for all user.

The Semantic Network and the SPECIALIST Lexicon, and their associated lexical tools, are accessible on the Internet (see [6] and [3]), under open terms. To view the terms and conditions for use of the Semantic Network and of the SPECIALIST Lexicon and Lexical Tools, see [5] and [4].

Since the Metathesaurus includes vocabulary content produced by many different copyright holders as well as the substantial content produced by NLM, it is necessary to establish a license agreement to use the Metathesaurus. This license agreement is set up via the web. After the license agreement is in place, the content of the Metathesaurus may be used under some conditions, available at [2].

2.1.1 UMLS components

The UMLS is made up of three main knowledge components: the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon. It also includes several tools that facilitate their use, including the MetamorphoSys, the install and customization program.

- Metathesaurus is the main part of UMLS, containing a collection of medical concepts and terms of the various vocabularies and the relationships between them. It represents several kinds of relations like hierarchical relations (”a is a part of b” or ”c is

a kind of d"), or associative relations ("e is caused by f" or "g often occurs close to h"). The use of different names for the same concept or the same name for different concepts, are also faithfully represented in the Metathesaurus [7].

- The Semantic Network, which provides a categorization of all concepts existent in the Metathesaurus and the semantic relationships that can be assigned between these concepts.

- The SPECIALIST Lexicon, which can be applicable in Natural Language Processing, is intended to be a general English lexicon with many biomedical terms. It contains the syntactic, morphological, and orthographic information of each word and term records.

2.1.2 Metathesaurus RRF (Rich Release Format) and ORF (Original Release Format) files

There are two different relational formats: the Riche Release Format (RRF) and the Original Release Format(ORF), available in the MetamorphoSys.

2.1.3 Structure of data

UMLS provides scripts to load the data on the RRF or ORF files into a database. The database server version recommend is MySQL Server 5.0.

For our purpose, we only need 8 tables. Those tables are described in Table 2.1 and their relationship is showed in Figure 2.1. There are also reference to the fields used from each table, which are described in Table 2.2

2.2 MEDLINE - Medical Literature Analysis and Retrieval System Online

MEDLINE is a bibliographic database owned by the National Library of Medicine, U.S.A. It contains over 17 million references to journal articles in life sciences and biomedical information. New citations are added Tuesday through Saturday.

The access to the database is free via the PubMed interface, or by downloading the XML files, after completing and submitting the License Request Form, available at [1].

About 48% of the citations added during 1995-2003 are for cited articles published in the U.S., about 88% are published in English, and about 76% have English abstracts written by authors of the articles. The Appendix A shows a MEDLINE citations. Not all

Table	**Used fields**	**Description**
MRCOC	CUI1, AUI1, CUI2, AUI2, SAB, COT, COF	Co-occurrence os concepts (concepts that occur together in the same entries)
MRCONSO	CUI, LUI, SUI, AUI, STR	Strings or concept names in the Metathesaurus
MRDEF	CUI, AUI, DEF	Definitions of atoms (occurrence of each string in a source vocabulary)
MRHIER	CUI, AUI, PAUI, RELA, PTR	Representation of all hierarchies present in the Metathesaurus
MRREL	CUI1, AUI1, CUI2, AUI2, REL, RUI	Representation of the relationships between the concepts known to the Metathesaurus
MRXW_ENG	CUI, LUI, SUI, WD	Words found in each unique Metathesaurus string
MRXNS_ENG	CUI, LUI, SUI, NWD	Normalized strings[1] found in each unique Metathesaurus string (only for English)
MRXNW_ENG	CUI, LUI, SUI, NSTR	Normalized words[2] found in each unique Metathesaurus string (only for English)

Table 2.1: Tables used from UMLS

the elements showed (*DateCreated*, *ArticleTitle*, etc.) are available for all the citations, like the *AbstractText*.

Since the objective was to gather a set of documents, a corpus, it was only necessary to keep the abstract text (text delimited by the <AbstractText> tag. It was also kept the title of the article (<ArticleTitle> tag) and the name of the file were it belongs (medline09nXXXX.xml, were XXXX is a number from 0000 to 0593).

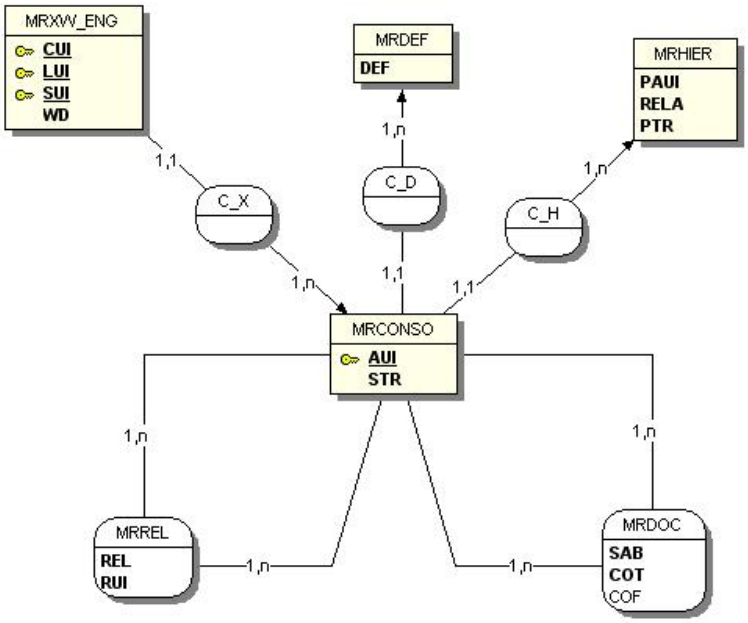

Figure 2.1: Entity-Relationship Diagram

Field	**Description**
AUI	Unique Atom Identifier
CUI	Unique Concept Identifier
SUI	Unique String Identifier
LUI	Unique Term Identifier
WD	Word
NWD	Normalized Word
STR	String
NSTR	Normalized String
DEF	Definition of atom
PAUI	Unique identifier of atom's immediate parent
RELA	Relationship of atom to its immediate parent
RUI	Unique Relationship Identifier
REL	Relationship of second concept or atom to first concept or atom
PTR	Path to the top of the hierarchical context from this atom, represented as a list of AUIs, separated by periods (.) The first one in the list is top of the hierarchy; the last one in the list is the immediate parent of the atom, which also appears as the value of PAUI.
SAB	Abbreviated source name
COT	Type of co-occurrence
COF	Frequency of co-occurrence

Table 2.2: UMLS database used fields

Chapter 3

Ontology Learning

3.1 Ontology

In [12], Tom Gruber defines *Ontology* as "*an explicit specification of a conceptualization.*".

For knowledge-based systems, it can only be represented what "exists", creating a set of objects. This set of objects, and the relationships between them, are shown in the representational vocabulary with which a knowledge-based program represents knowledge. Therefore, a ontology can be described by defining a set os representational terms. In this ontology, definitions associates the entities (classes, terms, relations, ...) with text describing the meaning of each entity.

In the Artificial Intelligence field, Ontology was initially defined by Neches et al. ([19]): "*An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary.*"

In [11], Pérez and Corcho identify the main components use in Ontologies:

1. **Concepts** are used in a broad sense and can be abstract or concret, elementary or composite, real or fictitious. A concept can be anything that was said about something and the description of a task, function, action, etc.

2. **Taxonomies** are used to organize ontological knowledge in the domain using generalisation/specialisation relationships.

3. **Relations** represent the connection between concepts of the domain and are formally defined as any subset of a product of n sets, this is, $R = C_1 \times C_2 \times C_3 \times \ldots \times C_n$ where C_1, C_2, C_3, ..., C_n represent concepts.

4. **Functions** are considered as a special kind of relations where the value of the last argument is unique for a list of values of the n-1 preceding arguments.

5. **Axioms** are included in an ontology to constrain its information (Example: a square is a rectangle which has four equal sides), verifying its correctness or deducting new information.

6. **Instances/Individuals/Facts/Claims** are used to represent elements in the domain. Instances represent elements of a given concept. Facts represent a relation which holds between elements. Individuals refer to any element in the domain which is not a class. Claims represent assertions of a fact by an instance.

Their are many types of relationships that can be established between terms. In [18], Miller lists the relations present in WordNet, which can be present in any Ontology:

- **Synonym** - terms having the same or nearly the same meaning. Example: *Ache* and *Pain*.
- **Antonym** - terms having a opposite meaning. Example: *Healthy* and *Unwell*.
- **Hyponym** - a term that is less general then another. Example: *Aorta* and *Arteries*.
- **Hypernym** - a term that is more general then another. Example: *Blood Vessels* and *Arteries*.
- **Meronym** - a term that is a part of another term. Example: *Eyes* and *Head* - Eyes is a part of head.
- **Holonym** - a term which names the whole of another term. Example: *Head* and *Nose* - Head has a nose.
- **Troponym** - same as hyponym, but it's applied to verbs, this is a verb that indicates a manner of doing something by replacing a verb of a more generalized meaning. Example *to march* and *to walk*.
- **Entailment** - encodes the relations between verbs.

In [21], van Heijst et al. classifies ontologies according to two dimensions: (1) the amount and type of structure of the conceptualization, and (2) the subject of the conceptualization.

In (1), they distinguish three categories:

1. **Terminological ontologies** which specify the terms that are used to represent knowledge in the domain of discourse. Example: UMLS.

2. **Information ontologies** which specify the record structure of databases.

3. **Knowledge modelling ontologies** specify conceptualizations of the knowledge. This ontologies usually have a richer internal structure than information ontologies. Also, they are often tuned to a particular use of the knowledge that they describe.

In (2), there are four categories:

1. **Application ontologies** which contains all the definitions that are needed to model the knowledge required for a particular application. They are a mix of concepts taken from domain ontologies and generic ontologies.

2. **Domain ontologies** express conceptualizations that are specific for particular domains.

3. **Generic ontologies** are similar to domain ontologies, but the concepts that they define are considered to be generic.

4. **Representation ontologies** provide a representational framework without making claims about the world. They provide primitives used to describe domain ontologies and generic ontologies.

Not agreeing with van Heijst, Guarino ([13]) suggest that a distinction can be made based on (1) the degree of detail used to characterize a conceptualization and (2) the subject of the conceptualization.

In (1), Guarino distinguish two kinds of ontologies:

1. **Documenting ontologies**, or off-line ontologies, are very rich in details and gets closer to specifying the intended conceptualization.

2. **Shareable ontologies**, or on-line ontologies, are very simple ontologies, which are developed with particular inferences, previously agreed by the user on the underlying conceptualization.

In (2), ontologies are distinguish in:

1. **Domain ontologies** express conceptualizations that are specific for particular domains.

2. **Method ontologies** express conceptualizations of a specific task or method, including all the concepts necessary to describe the inferential process, from the very abstract concepts to the more specialized concepts.

3. **Application ontology** which contains the definitions specific to a particular application.

3.2 Building an Ontology

Making an Ontology is a complex and slow process. Preparing a computer to do so without supervision is an even more complex and slow process.

First of all, it is necessary to prepare a base, so that we could compare and, in the future, evaluate the results of our work. To do so, we started by searching for a base of knowledge in the medical domain, the UMLS (see section 2.1). The next step was to select the words we would use. This selection was made for four domains in Medicine: (1) Cardiovascular System; (2) Digestive System; (3) Respiratory System and (4) Nervous System. For each of these domains/terms, it was chosen five hyponyms (if they exist), this is, five words that are directly below in the hierarchy present in UMLS. The process was repeated until we had six levels of generality, as showed in Appendix B.

Then, we processed the XML files from MEDLINE, saving the title, abstract text and the name of the file were the article was, as described in section 2.2.

3.3 Similarity Matrix

The next step is to calculate the Similarity Matrix. For this, we need to calculate the asymmetry between the terms, using Asymmetric Association Measures. From all the existing, we enumerate only seven, from which we used the Confidence measure during our work.

3.3.1 Asymmetric Association Measures

As stated by [16], asymmetry is important in Natural Language Processing, therefore in all of is strands, like ontology building. The asymmetry allows us to induce oriented associations between terms, like *dentition* and *mouth*. When someone hears *dentition*, they may think about *mouth*, but when hearing *mouth* more common elements will come to mind, like *tongue* or *teeth*. In this case, there is an oriented associations between dentition and mouth (dentition $\longrightarrow$ mouth) which indicates that dentition attracts more mouth, than mouth attracts dentition, this is, mouth is more likely to be a more general term than dentition.

Therefore, we can use asymmetric association measures, presented below, to measure the degree of attractiveness between to terms, x and y, where $f(x, y)$, $P(x)$, $P(x, y)$ and N are respectively the frequency function, the marginal probability function, the joint probability function, the number of documents in the corpus.

$$\text{Braun-Blanquet} = \frac{f(x,y)}{\max(f(x,y) + f(x,\overline{y}), f(x,y) + f(\overline{x},y))} \tag{3.1}$$

$$\text{J measure} = \max \left[\begin{array}{l} P(x,y)\log\frac{P(y|x)}{P(y)} + P(x,\overline{y})\log\frac{P(\overline{y}|x)}{P(\overline{y})}, \\ \\ P(x,y)\log\frac{P(x|y)}{P(x)} + P(\overline{x},y)\log\frac{P(\overline{x}|y)}{P(\overline{x})} \end{array} \right] \tag{3.2}$$

$$\text{Confidence} = \max\left[P(x|y), P(y|x)\right] \tag{3.3}$$

$$\text{Laplace} = \max\left[\frac{N.P(x,y)+1}{N.P(x)+2}, \frac{N.P(x,y)+1}{N.P(y)+2} \right] \tag{3.4}$$

$$\text{Conviction} = \max\left[\frac{P(x)P(\overline{y})}{P(x,\overline{y})}, \frac{P(\overline{x})P(y)}{P(\overline{x},y)} \right] \tag{3.5}$$

$$\text{Certainty Factor} = \max\left[\frac{P(y|x)-P(y)}{1-P(y)}, \frac{P(x|y)-P(x)}{1-P(x)} \right] \tag{3.6}$$

$$\text{Added Value} = \max\left[P(y|x)-P(y), P(x|y)-P(x) \right] \tag{3.7}$$

All these measures evaluate the maximum value between the attraction of x upon y and the attraction of y upon x to show the asymmetry between x and y. The maximum value decides de direction of the general-specific association, this is, ($x \longrightarrow y$) or ($y \longrightarrow x$).

Although all the measures were calculated and saved,we used only the Confidence measure to describe our work. In the future, it will be tested with the other measures, so that we can compare the results and choose the one with better results.

3.3.2 Calculating the Similarity Matrix

The final values of the Asymmetric Association Measures were calculated by steps, this is, it was first calculated the values of each part ($f(x,y)$ e $f(y,x)$), which was saved in a matrix (see table 3.1). Then, we calculated the maximum value ($max[f(x,y), f(y,x)]$), saving it in the matrix by putting 0 (zero) in the minimum value, originating the Similarity Matrix (see table 3.2).

	W_1	W_2	$\cdots$	W_n
W_1	1	$sim(W_1, W_2)$	$\cdots$	$sim(W_1, W_n)$
W_2	$sim(W_2, W_1)$	1	$\cdots$	$sim(W_2, W_n)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
W_n	$sim(W_n, W_1)$	1	$\cdots$	1

Table 3.1: Values of the Asymmetric Association Measures

	W_1	W_2	$\cdots$	W_n
W_1	1	$sim(W_1, W_2)$	$\cdots$	0
W_2	0	1	$\cdots$	$sim(W_2, W_n)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
W_n	$sim(W_n, W_1)$	0	$\cdots$	1

Table 3.2: Similarity Matrix

With the previous example, we obtained the following results:

(a) Confidence(mouth, dentition) = 0.00760456273764259

(b) Confidence(dentition, mouth) = 0.025974025974026

The maximum value is Confidence(dentition, mouth), which implies that mouth is more general than dentition, this is, (dentition $\longrightarrow$ mouth), like we stated above.

3.4 TextRank Algorithm

The use of a TextRank algorithm based on a graph is a way to decide the importance of a vertex (in our case, a term) within a graph, based on the information recursively drawn from the graph. With the use of a graph-based ranking algorithm, we intend to show that more general terms will be more likely to have incoming associations as they will be associated to many specific terms. In the other and, more specific terms will have more outgoing associations than incoming associations as they attract general terms (See Figure 3.1). Therefore, a graph-based ranking algorithms should give more strength to general terms than specific ones, ranking the term from general to specific.

Before using a graph-based ranking algorithms, we need to build a direct graph. As stated above (See section 3.3.1), if a term x attracts more another term y, than y attracts x, it will be created an edge between x and y as follows ($x \longrightarrow y$), so that we can give more credit to general words.

Formally, a directed graph $G = (V, E)$ can be defined with the set of vertices V (in our case, a set of terms) and a set of edges E, where E is a subset of $V \times V$ (in our case,

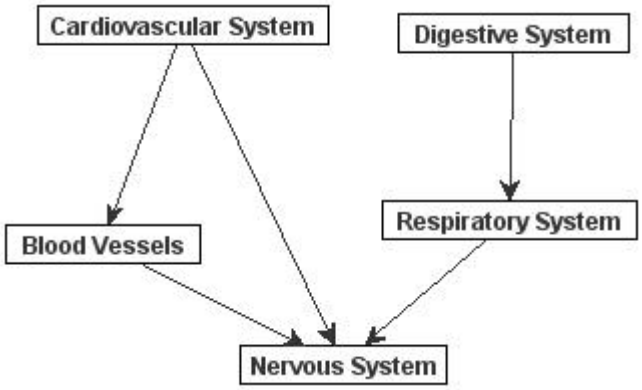

Figure 3.1: Directed Graph based on the Confidence measure.

a similarity matrix calculated by the asymmetric association measure value between two terms - See table 3.2).

In Figure 3.1, we show the directed graph obtained by using the first terms selected from UMLS, this is, the set V={"Cardiovascular System", "Digestive System", "Respiratory System", "Nervous System", "Blood Vessels"}. The weight of the edges have been calculated by the Confidence Association Measure (Equation 3.3) based on the corpus[1] counts. The Joint Probability between two terms, $P(x, y)$, is calculated by the number of documents where the terms x and y appear together divided by the total number of documents in the corpus. The Marginal Probabilities $P(x)$ and $P(y)$ are calculated by the same process.

By using a graph-based algorithm, we aim to produce an ordered list of terms from the most general (with the highest value) to the most specific (with the lowest value).

The algorithm used is the one proposed by [17], which can be used by Weighted and Unweighted Direct Graphs.

For a given vertex V_i, let $In(V_i)$ be a set of vertices that point to it, and $Out(V_i)$ a set of vertices that vertex V_i points to. The score of a vertex V_i is defined by:

$$S(V_i) = (1 - d) + d \times \sum_{V_j \in In(V_i)} \frac{1}{|Out(V_j)|} \times S(V_j) \tag{3.8}$$

where d is a damping factor that can be set between 0 and 1 (usually set to 0.85), which has the role of integrating into the model the probability of jumping from a given vertex to another random vertex in the graph.

To take into account the weight of the edges, it is introduced a new formula:

[1] set of texts extracted from the MEDLINE XML files

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} \times WS(V_j) \tag{3.9}$$

In our work, we use the Weight Direct Graph to calculate the terms rank. After running the algorithm, a score is associated to each vertex, representing the importance of the vertex within the graph. In Table 3.3, we show the result for the terms included in the set V described above. The results for all the terms selected from UMLS is shown in Appendix C.

Pos.	AUI	Term	Rank Value
3.	A0091423	Nervous System	3.4234311255891776
8.	A0031694	Blood Vessels	0.9797096311214579
51.	A0035256	Cardiovascular System	0.19773753019147863
62.	A0111426	Respiratory System	0.17384242429857602
74.	A0048766	Digestive System	0.15000000000000002

Table 3.3: Partial results of WordRank

Although the results do not correspond to the reality - Cardiovascular System, Respiratory System, Digestive System should be higher in the rank, for they are more general, we can consider them satisfactory. The position of these terms can be related to the fact that, usually, when talking about them, the term is not used, this is, we can talk about ”Blood Vessels” without talking about ”Cardiovascular System”. Therefore, more general words can be considered more specific.

3.5 Determination of clusters of words

Cluster analysis or clustering is a filed of research that belongs to data analysis and machine learning major domains.

Clustering divides a set of data into groups whose members are similar to each other. Each object (called ”term” in the following), is assigned to the group it is more similar to. For instance, ”Hospital” and ”Doctor’s office” are similar for booth represent a place were medicine is practiced, therefore, they can be placed in the same group.

The main objective in clustering is placing similar terms in the same group and dissimilar words in different groups, which are called clusters.

Several ways of clustering have been explored, originating some types of algorithm. There are the *hard-clustering* techniques, which assigns each term to a single cluster. Inversely, in fuzzy-clustering methods, terms can belong to more than one cluster, and associ-

ated with each term is a set of values which indicate the strength of the association between that term and a particular cluster.

In our work, we tried to group the terms by degree of generality. For that, we used the *K*-Means algorithm (See subsection 3.5.1), which allows us to define the number of cluster (k), and the PoBOC algorithm (See subsection 3.5.2), that defines the better clusters for the data set, without defining the number of clusters *a priori*.

Then, we calculate the degree of generality by using the values returned by the TextRank algorithm, so that we could have the clusters ranked by degree of generality.

3.5.1 *K*-Means

K-Means algorithm was introduced by MacQueen in 1967, and is one of the most common clustering algorithm that groups data with similar characteristics together. The data in the same cluster will have similar characteristics, which are dissimilar from the data in other cluster.

The steps of the algorithm are:

1. Introducing the number of clusters (k[2]) and the data set to be clustered.

2. Choose the k initial clusters, by randomly choosing k rows of data from the data set.

3. Each row of data (each term) will be assigned to one of the initial clusters. After evaluating the similarity between the row of data and the initial clusters, the row of data is assigned to the cluster it is most similar to (the nearest cluster). Measures like Distance Measure and Asymmetric Association Measures can be use to calculate the similarity. In our case, we used the Confidence Measure.

4. The Arithmetic Mean, or center, of each cluster is re-calculated (the initial Arithmetic Mean is the set of values of the data row chosen for the initial cluster). The Arithmetic Mean of a cluster is the mean of all the individual data rows (terms) in the cluster. For example, with the set of values for the Confidence and Laplace measures ({Confidence, Laplace}), if a cluster contains two data rows with the measurements for $mouth = \{0.0547, 1.2541\}$ and $dentition = \{0.258741, 0.0125454\}$, then the arithmetic means is represented by $Cl_{mean} = \{Confidence_{mean}, Laplace_{mean}\}$, where $Confidence_{mean} = (0.0547 + 0.258741)/2$ and $Laplace_{mean} = (1.2541 + 0.0125454)/2$. The arithmetic means of this cluster is $\{0.1567205, 0.6333227\}$.

[2]In our case, we used $k = 6$, so that we could compare with the UMLS structure.

5. Each row of data (each term) will be re-assigned to one of the new clusters initiated by the Arithmetic Mean.

6. The steps **4** and **5** are repeated until **stable clusters** are formed, this is, until the repetition of the k-means algorithm does not create different or new clusters as the cluster center or Arithmetic Mean of the formed cluster is the same as the old cluster center.

Although the procedure will terminate, it does not necessarily find the most optimal clusters and, as the algorithm is sensitive to the initial randomly selected cluster center, the result can be different every time we run the algorithm.

The implementation of the k-means algorithm is the one developed by Cleuziou ([10]), the OKM algorithm (Overlapping *k*-Means).

In his approach, Cleuziou defines an objective criterion: given a set of data vectors $\chi = \{x_i\}_{i=1}^{n}$ with $x_i \in \mathbf{R}^p$, the goal of the OKM algorithm (Overlapping *k*-means) is to find a *k*-way coverage $\{\pi_c\}_{c=1}^{k}$ of the date (where π_c represents the c^{th} cluster) such that the following objective is minimized:

$$\jmath(\{\pi_c\}_{c=1}^{k}) = \sum_{x_i \in \chi} \|x_i - \phi(x_i)\|^2 \tag{3.10}$$

Each x_i belongs at least to one cluster, since $\{\pi_c\}_{c=1}^{k}$ is a coverage, which is such that $\cup_{c=1}^{k}\pi_c = \chi$. This way, $\phi(x_i)$ indicates the "image" of x_i defined by combination of the prototypes (m_c) for the clusters x_i belong to:

$$\phi(x_i) = \frac{\sum_{A_i} m_c}{|A_i|} \tag{3.11}$$

where A_i denotes the set of assignment for $x_i : \{m_c | x_i \in \pi_i\}$.

The Figure 3.2 gives an overview of the OKM algorithm.

The OKM implementation allows a term to belong to more than one clusters, but, in our work, we are only interested (for now) to have a term belonging to only one cluster. Therefore, when executing OKM we use the option $-o = 0$ which indicates that the algorithm should use hard clustering. With $-o = 1$, the algorithms use overlapping clusters, this is, a term can be assigned to more than one cluster.

The Appendix D lists the clusters resulting from the *K*-Means algorithm, after they were rank by the degree of generality (See section 3.6).

OKM($\mathcal{X}$,t_{max},ϵ)

Input: $\mathcal{X}$: a set of data vectors in $\mathbb{R}^p$, t_{max}: optional maximum number of iterations, ϵ: optional threshold on the objective

output: $\{\pi_c\}_{c=1}^k$: final coverage of the points

1. Draw randomly k initial cluster prototypes $\{m_c^{(0)}\}_{c=1}^k$ in $\mathbb{R}^p$ or $\mathcal{X}$.

2. For each $x_i \in \mathcal{X}$ compute the assignments $A_i^{(0)} = \text{ASSIGN}(x_i, \{m_c^{(0)}\}_{c=1}^k)$ and derive the initial coverage $\{\pi_c^{(0)}\}_{c=1}^k$ such that

$$\pi_c^{(0)} = \{x_i | m_c^{(0)} \in A_i^{(0)}\}$$

3. Set $t = 0$.

4. For each cluster $\pi_c^{(t)}$ successively, compute the new prototype

$$m_c^{(t+1)} = \text{PROTOTYPE}(\pi_c^{(t)})$$

5. For each $x_i \in \mathcal{X}$ compute the assignments $A_i^{(t+1)} = \text{ASSIGN}(x_i, \{m_c^{(t+1)}\}_{c=1}^k, A_i^{(t)})$, and derive the new coverage $\{\pi_c^{(t+1)}\}_{c=1}^k$.

6. If not converged or $t_{max} > t$ or $\mathcal{J}(\{\pi_c^{(t)}\}) - \mathcal{J}(\{\pi_c^{(t+1)}\}) > \epsilon$, set $t = t+1$ and go to Step 4; Otherwise, stop and output final clusters $\{\pi_c^{(t+1)}\}_{c=1}^k$.

Figure 3.2: OKM overview.

3.5.2 PoBOC

The algorithm PoBOC (Pole-Based Overlapping Clustering) is a method which can be viewed as a compromise between *hard* and *fuzzy-clustering* approaches [9].

One of the advantages of PoBOC, with regard to methods based on centroids is that each group is represented by a pole, this is, a set of objects, instead of only one object, which is a less specific representation of a group.

The algorithm PoBOC takes the previous calculated similarity matrix (See subsection 3.3.2) as input and builds in a hierarchy of concepts in which an object may belong to several concepts, this is, produces a set of overlapping clusters hierarchically organized.

From the similarity matrix SM defined on $X \times X$ over set of terms $X = x_1, x_2, \ldots, x_n$, the algorithm executes the following steps:

1. construction of the similarity graph $G_S(X, V)$, where V is the set of edges between

the terms;

2. extraction of complete sub-graphs from $G_S(X, V)$, the "poles";

3. multi-assignment of the terms to the poles;

4. hierarchical organization of the obtained groups.

In the first step, X is already defined as the set of terms obtained from UMLS, and is the base of the similarity graph denoted by $G_S(X, V)$. V is the set of edges between the terms. We can say there is an edges between two terms x_i and x_j if x_i belongs to the neighborhood of x_j and vice-versa, which can be defined by:

$$s(x_i, x_j) \geq \max\{\frac{1}{n}\sum_{x_k \in X} s(x_i, x_k), \frac{1}{n}\sum_{x_k \in X} s(x_j, x_k)\} \tag{3.12}$$

With this definition, exists an edge between x_i and x_j if their similarity is greater than both the average similarity between x_i and the whole set of objects and the average similarity between x_j and the whole set of objects. This allows to take into consideration the density around each object.

The second step uses two heuristics to extract the poles from the similarity graph. A pole P_k is a subset of X such that the sub-graph $G_S(P_k, V(P_k))$ is a clique-graph, this is, $\forall x_i \in P_k$, $\forall x_j \in P_k$, $(x_i, x_j) \in V(P_k))$ with $V(P_k)$ the subset of edges associated with P_k.

The Poles are derived from the similarity graph $G_S(X, V)$ by selecting a particular point and adding connected vertices to the pole in construction. The first vertex select (x^1) his the one with the lower average similarity with the other objects, among the set of vertices having at least one connected vertex. The set of vertices having at least one connect vertex, E, is include in X ($E \subset X$). x^1 is defined by:

$$x^1 = \arg\min_{x_i \in E} \frac{1}{|X|}\sum_{x_j \in X} s(x_i, x_j) \tag{3.13}$$

The next vertices $\{x^2, x^3, \ldots, x^m\}$ are chosen in order to reduce the similarity with the poles previously built:

$$x^k = \arg\min_{x_i \in E} \frac{1}{k-1}\sum_{m=1,\ldots,k-1} \frac{1}{|P_m|}\sum_{x_j \in P_m} s(x_i, x_j) \tag{3.14}$$

The process stops when the sum in the previous equation is greater than the average similarity of the whole set of objects. The number of poles is determined by 3.14, and correspond to the number of clusters.

The next step is the multi-assignment of the terms to one or several poles, among the set of poles $P = \{P_1, ..., P_l\}$. This step plays an important role in the construction of overlapping clusters in PoBOC. The advantage of assigning an object to several clusters is that way we can have a word with several senses in different clusters. Each cluster should represent optimally a sense of the word. In other soft-clustering algorithms, the assignment method is often based on a global threshold applied on a membership matrix obtained with a fuzzy-clustering method.

Cleziou et al. ([9]) proposes a new approach for the multi-assignment of objects to the poles based on the relative similarity between objects and poles. Let $X = x_1, \ldots, x_n$ be the set of objects, let $P = P_1, \ldots, P_n$ be the set of poles and U be the membership matrix on $P \times X$ defined by:

$$u(P_i, x_j) = \frac{1}{|P_i|} \sum_{x_k \in P_i} s(x_j, x_k) \tag{3.15}$$

Given an object x_j, $P_{j,1}$ is the most similar pole for x_j ($P_{j,1} = \arg\max_{P_1 \in P} u(P_i, x_j)$), $P_{j,2}$ the second most similar pole for x_j, and so on, $P_{j,l}$ is the least similar pole for x_j. The following condition $Assign(x_j, P_{j,k})$ is used to test whether the object x_j is assigned to the pole $P_{j,k}$: $Assign(x_j, P_{j,k})$ iff one of the following properties is satisfied:

i) $k = 1$,

ii) $1 < k < l, u(P_{j,k}, x_j) \geq \frac{u(P_{j,k-1}, x_j) + u(P_{j,k+1}, x_j)}{2}$ and $\forall k', Assign(x_j, P_{j,k'})$

where l is the number of resulting overlapping clusters.

For each object, the set of poles is ordered by his average similarity with the object. The property i) allows to assign each object to at least one pole (the most similar) and ii) allows to assign an object x_j to a pole $P_{j,k}$ by considering the similarity with the previous pole ($P_{j,k-1}$) and the next pole ($P_{j,k+1}$) w.r.t. the order associated to x_j. The assignment is based on the relative position (similarity) of the object with regards to the poles previously obtained.

The last step is the hierarchical organization of groups, which allows to control the number of final clusters obtained from the set of groups. This will be used so that it could be found a better characterization of the groups, and then to get a better set of clusters.

To get a hierarchical organization, Cleziou et al. propose to apply the hierarchical agglomerative clustering method "single-link", described by Kain, Murty and Flynn ([8]). Starting with the groups previously built $C = C_1, \ldots, C_l$ where $C_i = x_j \ assigned \ to \ P_i$. Since the similarity matrix is normalized, we have $\forall x_i \in X, s(x_i, x_i) = 1$ and the similarity between two groups is defined by:

$$sim(C_k, C_m) = \frac{1}{|C_k| \cdot |C_m|} \sum_{x_i \in C_k} \sum_{x_j \in C_m} s(x_i, x_j) \tag{3.16}$$

The organization of clusters is built as follows: the two most similar groups are agglomerated and this process is repeated until we get only one group. This organization is represented by a binary tree where the leaves correspond to the initial set of groups.

After the execution of PoBOC in our set of terms, the algorithm returns a set of clusters, from which we only use the ones with more than one term.

The Appendix E lists the clusters resulting from the PoBOC algorithm, after they were rank by the degree of generality (See section 3.6).

3.6 Cluster Ranking

As the base of comparison of our work is ranked by degree of generality and this does not happens with the cluster returned by the algorithm, it is necessary to rank the clusters. This is done using the value obtained with the TextRank algorithm (See section 3.4 and Appendix C).

For each cluster we calculate the average of the TextRank value of the terms assign to it.

For example, using the following clusters, obtained by the PoBOC algorithm:

```
GenCluster 2                          GenCluster 4
-----------------------------------   -----------------
(0,1) ---> Cardiovascular System      (0, 5) ---> Blood Vessels
(1,2) ---> Digestive System           (1,46) ---> Oropharynx
(2,3) ---> Respiratory System         (2,93) ---> Geniculate Ganglion
(3,4) ---> Nervous System
```

and the TextRank values, with the values rounded to five decimal places:

AUI	Term	Rank Value
A0091423	Nervous System	3.30660
A0031694	Blood Vessels	0.92931
A0095412	Oropharynx	0.27398
A0035256	Cardiovascular System	0.19745
A0111426	Respiratory System	0.17371
A0048766	Digestive System	0.15000
A0062894	Geniculate Ganglion	0.15000

we obtain the following values:

$$RankValue(GenCluster2) = \frac{3.30660 + 0.19745 + 0.17371 + 0.15000}{4} = 0.95694$$

and

$$RankValue(GenCluster4) = \frac{0.92931 + 0.27398 + 0.15000}{3} = 0.45110$$

Comparing the two values, the higher value is $RankValue(GenCluster4)$, therefore, "GenCluster 2" is more general than "GenCluster 4", appearing before in the cluster ranked list, as we can see in Appendix E.

Appendix D shows one of the results of the *k*-means algorithm, as it shows a different result each time the algorithm is run, after ranking the clusters.

3.7 Domain Clustering

The next step is to determinate domain clusters within the generality clusters obtained.

For this, we use the Symmetric Conditional Probability (SCP) measure, which calculates numerically the similarity between terms, based on the co-occurrence frequency acquired from the corpus.

The SCP measure is defined by:

$$SCP(w_1, w_2) = p(w_1|w_2) \times p(w_2|w_1) = \frac{p(w_1, w_2)^2}{p(w_1) \times p(w_2)} \tag{3.17}$$

This process is done by:

1. Calculate the SCP measure between the terms within each Cluster.
2. Build similarity matrix for each cluster,using the SCP values.
3. Re-use the clustering algorithm in the clusters returned from the first use (See section 3.5.2).
4. Rank the clusters (See section 3.6).

The domain cluster obtained for the first generality cluster (GenCluster 1), with PoBOC algorithm, is:

```
DomCluster 1
-----------------------
   (  0,   6) ---->  Nose
   (  1,  13) ---->  Salivary Glands
   (  2,  15) ---->  Maxillary Sinus
   (  3,  17) ---->  Submandibular Gland

DomCluster 2
-----------------------
   (  0,   3) ---->  Pancreas
   (  1,   7) ---->  Pleura
```

```
(  2,  14) ---> Goblet Cells
(  3,  16) ---> Colon

DomCluster 3
--------------------
(  0,   4) ---> Larynx
(  1,   5) ---> Lung
(  2,  11) ---> Hypopharynx
(  3,  12) ---> Nasopharynx

DomCluster 4
--------------------
(  0,   1) ---> Gastrointestinal Tract
(  1,   2) ---> Liver
(  2,   8) ---> Lower Gastrointestinal Tract
```

Note that, as referred in 3.5.2, we only use the clusters with more than one terms. This explains why the terms "Biliary Tract", "Pulmonary Alveoli" and "Nasal Mucosa", existing in PoBOc's 'GenCluster 1', do not appear in this domain clusters.

3.8 Problems found

Building an language-independent Ontology without supervision is a difficult process, during which several problems can be found. During the elaboration os this work we encountered the following problems:

1. The small size of the documents in the corpus. As MEDLINE only keeps references to journal articles,only the abstract of some articles are in those references.

2. As we intend to keep the system language-independent, the variations of the terms are treated distinctly. For example *lung* and is plural, *lungs*, can be treated as the same term in a language-dependent system, but, on a language-independent system, they era completely different terms, as the rules to inflict terms variations are language-dependent.

3. The more general terms, like *Cardiovascular System*, are rarely used, therefore, they can be wrongly classified. So, more general words can be considered more specific, and vice-versa.

4. The size of the compound terms, this is, the number of words that defines the compound term, complicates the process, as the bigger the term is, more difficult it is to find it in the corpus.

Chapter 4

Related Work

Oliveira ([20]) also aims to the automatic creation of a lexical Ontology, but for the Portuguese language.

The resulting Ontology of Oliveira's work is intended to be in the public domain and freely available for download, so it could be used by the Portuguese NLP community and other researchers that need Portuguese lexical knowledge in their work.

Lexical capabilities of NLP systems are weak, as there is intensive work involved in manually encoding lexical entries, which is impractical and undesirable. NLP tools should be used in order to automate part of the process, decreasing the manual input.

Oliveira talks about lexical database, where concepts are organized in a network and relate with other concepts by means of semantic relations, which are present in text.

As Oliveira's work is language-dependent, he can take advantage of the language characteristics, like textual patterns which can be used to identify semantic relations. He also intends to use machine readable dictionaries (MRDs) to acquire general knowledge and then corpora to complete specific lexical gaps.

Oliveira's choice on using MRDs is related to the fact that "*MRDs are highly structured, they are a substantial source of general lexical knowledge, and the "authorities" of word sense*". They are also used for word sense disambiguation (WSD). He intends to use PAPEL [1]as a base his our ontology, by analysing its structure and relations in order to improve the grammars, the extraction tools and, therefore, the quality of the relations. Besides PAPEL, he is also planning on using other MRDs, like the portuguese version of Wiktionary (http://pt.wiktionary.org), with the objective of improving is work.

Resembling our work, Oliveira will also extract relations from textual resources, like textual corpora, which should be viewed as the main source of domain-specific information

[1]a lexical resource for Portuguese, consisting of relations between terms, extracted after processing the definitions of a major general dictionary

([14]). Therefore, Oliveira will try to develop tools to extract relations from corpora, which will be used to enrich the main Ontology, our to build new ones based on the texts.

For extracting relations from corpora, Oliveira refers Hearst's work ([15]), which propose an automatic method to discover lexico-syntatic patterns, used for the acquisition of hyponyms. Other researchers used Hearst proposal, not only for the hyponyms, but also for other relations like meronymy or causality.

He will also try to adapt the tools used to extract relations from MRDs to textual resources, knowing that extraction from unrestricted text is a more difficult task, as this text is not structured, its vocabulary not controlled and may contain several features like metaphors and anaphora.

The main relation with our work is the extraction os relations from corpora, although Oliveira uses different methods, as we will be working for a specific language, Portuguese, taking advantage of its characteristics, as the rules of words derivations.

Chapter 5

Conclusions and Future Work

We can say that ontologies will play a crucial role in future knowledge-based systems if they will be designed in such a way as to minimize the effects of the interaction problem. Ontologies are described in the knowledge level, its objective is to characterize a conceptualization, trying to describe the interpretations of each object, in our case, of each term.

As described in section 3.1, an Ontology is defined as an explicit specification of a conceptualization, by Tom Gruber (1993).

Many types of relations, described by Pérez and Corcho (2000), can be extracted from corpora, but some of them are language-dependent, therefore, were not taken into count in our research, as we want to build a language-independent ontology.

Many work was done, and there much more to be done (See section 5.1), as building a language-independent ontology is a very complex and prolonged process.

Using features like asymmetric association measures and Symmetric Conditional Probability allowed us to define relations between terms, without being necessary the access to lexical resources already built. Although, some of these relations were not what we expected, mainly due to the problems described in section 3.8.

Those wrongly defined relations committed the result obtained in the rest of our approach. Still, the result can be considered satisfactory as we obtained the following domain clusters: $C_1 = {'GastrointestinalTract',' Liver,' LowerGastrointestinalTract'}$ and $C_2 = {'Larunx',' Lung',' Hypopharynx,' Nasopharynx'}$ which belong to 'Digestive System' and 'Respiratory System', respectively.

Therefore, we can say that, besides not concluding our work, we obtain satisfactory results, thus making our approach a point of start for developing language-independent ontologies.

The next section describes the future work, need to complete our approach, and some items that can improve our work.

5.1 Future work

As this thesis is incomplete, this is, the main objective was not achieved, there is some work to be done:

i) Find synonymy clusters for each domain clusters obtain in 3.7.

ii) Calculate the similarity between terms within the domain clusters, using word-context vectors, building a similarity matrix.

iii) Interconnect the clusters. This can be done by:

   a) Interconnect domain clusters;

   b) Interconnect synonym clusters.

iv) Test and evaluate our ontology.

After this, we can improve our ontology by:

i) Use word-context vectors in relations extraction.

ii) Use bigger documents, so we can improve the terms classification.

iii) Find a way of derive the score for compound terms, as it his difficult to calculate by the method used.

Bibliography

[1] *License Agreement for NLM Data*. http://www.nlm.nih.gov/databases/license/weblic/index.html.

[2] *License Agreement for Use of the UMLS Metathesaurus*. http://www.nlm.nih.gov/research/umls/metaa.html.

[3] *The SPECIALIST NLP Tools*. http://lexsrv3.nlm.nih.gov/SPECIALIST/index.html.

[4] *Terms and Conditions for Use of the SPECIALIST NLP Tools*. http://lexsrv3.nlm.nih.gov/SPECIALIST/docs/TermsAndConditions.html.

[5] *Terms and Conditions for Use of the UMLS Semantic Network*. http://semanticnetwork.nlm.nih.gov/TermsAndConditions/.

[6] *The UMLS Semantic Network*. http://semanticnetwork.nlm.nih.gov/.

[7] *Unified Medical Language System*. http://en.wikipedia.org/wiki/Unified_Medical_Language_System.

[8] M. N. Murty A. K. Jain and P. J. Flynn. Data clustering: a review. 2004.

[9] G. Cleuziou, L. Martin, and C. Vrain. Poboc: an overlapping clustering algorithm. application to rule-based classification and textual data. pages 440–444, August 22-27 2004.

[10] Guillaume Cleuziou. An extended version of the k-means method for overlapping clustering. In *ICPR*, pages 1–4. IEEE, 2008.

[11] Óscar Corcho and Asunción Gómez-Pérez. A roadmap to ontology specification languages. In *EKAW '00: Proceedings of the 12th European Workshop on Knowledge*

Acquisition, Modeling and Management, pages 80–96, London, UK, 2000. Springer-Verlag.

[12] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[13] Nicola Guarino. Understanding, building and using ontologies. *Int. J. Hum.-Comput. Stud.*, 46(2-3):293–310, 1997.

[14] M. Hearst. Automated discovery of wordnet relations. 2009.

[15] M. Hearst. Automatic acquisition of hyponyms from large text corpora. 2009.

[16] Lukas Michelbacher, Stefan Evert, and Hinrich Schütze. Asymmetric association measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2007*, 2007.

[17] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. In *Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 2004.

[18] George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[19] Robert Neches, Richard Fikes, Tim Finin, Tom Gruber, Ramesh Patil, Ted Senator, and William R. Swartout. Enabling technology for knowledge sharing. *AI Mag.*, 12(3):36–56, 1991.

[20] Hugo Gonçalo Oliveira. Ontology learning for portuguese. In *2nd Doctoral Symposium on Artificial Intelligence (SDIA)*, 2009.

[21] G. van Heijst, A. Th. Schreiber, and B. J. Wielinga. Using explicit ontologies in kbs development. *Int. J. Hum.-Comput. Stud.*, 46(2-3):183–292, 1997.

Appendix A

Example of a MEDLINE entry

```
<MedlineCitation Owner="NLM" Status="PubMed-not-MEDLINE">
  <PMID>16692695</PMID>
  <DateCreated>
    <Year>2006</Year>
    <Month>06</Month>
    <Day>01</Day>
  </DateCreated>
  <DateCompleted>
    <Year>2006</Year>
    <Month>06</Month>
    <Day>01</Day>
  </DateCompleted>
  <DateRevised>
    <Year>2008</Year>
    <Month>11</Month>
    <Day>20</Day>
  </DateRevised>
  <Article PubModel="Print">
    <Journal>
      <ISSN IssnType="Print">0065-9533</ISSN>
      <JournalIssue CitedMedium="Print">
        <Volume>23</Volume>
        <PubDate>
          <Year>1925</Year>
        </PubDate>
      </JournalIssue>
      <Title>Transactions of the American Ophthalmological Society</Title>
    </Journal>
    <ArticleTitle>Globular Masses on the Pupillary Margin in Acute
                  Circumscribed Chorioretinitis. Clinical and
                  Pathologic Study.
    </ArticleTitle>
    <Pagination>
      <MedlinePgn>106-20</MedlinePgn>
    </Pagination>
    <Abstract>
      <AbstractText>1. A group of cases is described in which inflammatory
```

```
                    lesions in the posterior segment of the eye were
                    associated with the deposit of cells on the margin
                    of the pupil and the posterior surface of the cornea.
                 2. The path of wandering cells in relation to localized
                    inflammatory lesions in the eye is discussed.
                 3. The pathologic findings of a case of chorioretinitis
                    at the optic disc are described.
                 4. An explanation is offered for the sector-formed defect
                    in the visual field of such cases.
        </AbstractText>
      </Abstract>
      <AuthorList CompleteYN="Y">
        <Author ValidYN="Y">
          <LastName>Friedenwald</LastName>
          <ForeName>H</ForeName>
          <Initials>H</Initials>
        </Author>
        <Author ValidYN="Y">
          <LastName>Friedenwald</LastName>
          <ForeName>J S</ForeName>
          <Initials>JS</Initials>
        </Author>
      </AuthorList>
      <Language>eng</Language>
      <PublicationTypeList>
        <PublicationType>Journal Article</PublicationType>
      </PublicationTypeList>
    </Article>
    <MedlineJournalInfo>
      <Country>United States</Country>
      <MedlineTA>Trans Am Ophthalmol Soc</MedlineTA>
      <NlmUniqueID>7506106</NlmUniqueID>
    </MedlineJournalInfo>
    <OtherID Source="NLM">PMC1316527</OtherID>
  </MedlineCitation>
```

Appendix B

Structure of terms selected from UMLS

Cardiovascular System
- Blood Vessels
  - Arteries
    - Aorta
      - Aorta, Abdominal
      - Sinus of Valsalva
      - Aorta, Thoracic
    - Arterioles
    - Axillary Artery
    - Basilar Artery
    - Brachial Artery
  - Endothelium, Vascular
    - Tunica Intima
    - Pericytes
  - Microcirculation
    - Arterioles
    - Capillaries
    - Venules
    - Arteriovenous Anastomosis
  - Muscle, Smooth, Vascular
    - Tunica Media
  - Retinal Vessels
    - Retinal Artery
    - Retinal Vein
- Blood–Air Barrier
- Blood–Brain Barrier
- Blood–Retinal Barrier
- Blood–Testis Barrier

Digestive System
- Biliary Tract
  - Bile Ducts
    - Bile Ducts, Intrahepatic
      - Bile Canaliculi

- Bile Ducts, Extrahepatic
  - Common Bile Duct
    - Ampulla Of Vater
      - Sphincter Of Oddi
    - Cystic Duct
    - Hepatic Duct, Common
- Gallbladder

- Gastrointestinal Tract
  - Intestines
    - Intestinal Mucosa
      - Paneth Cells
      - Goblet Cells
      - Enterocytes
    - Intestine, Large
      - Cecum
        - Appendix
      - Colon
        - Colon, Ascending
        - Colon, Transverse
        - Colon, Descending
        - Colon, Sigmoid
      - Rectum
      - Anal Canal
    - Intestine, Small
      - Duodenum
        - Brunner Glands
        - Ampulla Of Vater
          - Sphincter Of Oddi
      - Ileum
        - Ileocecal Valve
        - Meckel Diverticulum
      - Jejunum
  - Pharynx
  - Upper Gastrointestinal Tract
    - Duodenum
      - Brunner Glands
      - Ampulla Of Vater
        - Sphincter Of Oddi
    - Esophagus
      - Esophagogastric Junction
        - Esophageal Sphincter, Lower
      - Esophageal Sphincter, Upper
    - Stomach
      - Cardia
      - Esophagogastric Junction
        - Esophageal Sphincter, Lower
      - Gastric Fundus
      - Gastric Mucosa
        - Enterochromaffin Cells
        - Parietal Cells, Gastric
        - Gastrin-Secreting Cells
        - Somatostatin-Secreting Cells
        - Chief Cells, Gastric
      - Pyloric Antrum

```
        Lower Gastrointestinal Tract
          – Ileum
              Ileocecal Valve
              Meckel Diverticulum
          – Intestine, Large
              Cecum
                – Appendix
              Colon
                – Colon, Ascending
                – Colon, Transverse
                – Colon, Descending
                – Colon, Sigmoid
              Rectum
              Anal Canal
          – Jejunum
        Mouth
          – Dentition
          – Salivary Glands
              Parotid Gland
              Salivary Glands, Minor
              Sublingual Gland
              Submandibular Gland
              Salivary Ducts
          – Tongue
              Lingual Frenum
              Taste Buds
    – Liver
        Bile Ducts, Intrahepatic
          – Bile Canaliculi
    – Pancreas
        Islets of Langerhans
          – Glucagon–Secreting Cells
          – Insulin–Secreting Cells
          – Somatostatin–Secreting Cells
          – Pancreatic Polypeptide–Secreting Cells
        Pancreatic Ducts
        Pancreas, Exocrine

Respiratory System
    – Larynx
        Glottis
          – Vocal Cords
        Laryngeal Cartilages
          – Arytenoid Cartilage
          – Cricoid Cartilage
          – Epiglottis
          – Thyroid Cartilage
        Laryngeal Mucosa
          – Goblet Cells
        Laryngeal Muscles
    – Lung
        Bronchi
        Extravascular Lung Water
        Pulmonary Alveoli
```

- Blood–Air Barrier
- Nose
  - Nasal Bone
  - Nasal Cavity
  - Nasal Mucosa
    - Olfactory Mucosa
      - Olfactory Receptor Neurons
    - Goblet Cells
  - Nasal Septum
    - Vomeronasal Organ
  - Paranasal Sinuses
    - Ethmoid Sinus
    - Frontal Sinus
    - Maxillary Sinus
    - Sphenoid Sinus
- Pharynx
  - Hypopharynx
  - Nasopharynx
  - Pharyngeal Muscles
    - Esophageal Sphincter, Upper
  - Oropharynx
    - Tonsil
- Pleura

Nervous System
- Ganglia
  - Ganglia, Autonomic
    - Ganglia, Parasympathetic
    - Ganglia, Sympathetic
      - Stellate Ganglion
      - Superior Cervical Ganglion
  - Satellite Cells, Perineuronal
  - Ganglia, Sensory
    - Ganglia, Spinal
    - Geniculate Ganglion
    - Nodose Ganglion
    - Spiral Ganglion
    - Trigeminal Ganglion
  - Ganglia, Invertebrate
- Nerve Net
  - Ependyma
  - Neurilemma
- Nerve Tissue
- Neural Pathways
  - Afferent Pathways
    - Auditory Pathways
    - Olfactory Pathways
    - Spinothalamic Tracts
    - Visual Pathways
    - Visceral Afferents
  - Efferent Pathways
    - Extrapyramidal Tracts
    - Pyramidal Tracts
  - Medial Forebrain Bundle

```
        Internal Capsule
        Perforant Pathway
  – Neuroglia
        Astrocytes
        Oligodendroglia
           – Myelin Sheath
        Schwann Cells
           – Myelin Sheath
                Neurilemma
                Ranviers Nodes
        Satellite Cells, Perineuronal
        Microglia
```

Appendix C

Results of TextRank - Confidence measure

Pos.	AUI	Term	Rank Value
1.	A0080671	Liver	11.199297772879172
2.	A0081228	Lung	5.349729619640907
3.	A0091423	Nervous System	3.306601449017083
4.	A0040976	Colon	2.058363909611326
5.	A0026415	Arteries	2.057223946618571
6.	A0119769	Stomach	1.6630908765684063
7.	A0088216	Mouth	1.059277666093959
8.	A0097048	Pancreas	0.9391519363076951
9.	A0031694	Blood Vessels	0.929311170267446
10.	A0126767	Tongue	0.8945121514369179
11.	A0078314	Larynx	0.8672044692096181
12.	A0056154	Esophagus	0.7847461248756237
13.	A2783410	Gastrointestinal Tract	0.7838685894965458
14.	A0034558	Capillaries	0.7352137286077568
15.	A0025439	Aorta	0.7101394780688466
16.	A0061914	Ganglia	0.7076406103397542
17.	A0051844	Duodenum	0.6364784462809864
18.	A0090234	Nasal Cavity	0.6192382063057026
19.	A0072068	Ileum	0.5804437631179689
20.	A0027374	Astrocytes	0.5185591679998053
21.	A0113722	Salivary Glands	0.4819805274060969
22.	A0093098	Nose	0.4779661328475734
23.	A0061720	Gallbladder	0.4161215490182107
24.	A0086379	Microcirculation	0.39845481665412685
25.	A0083644	Maxillary Sinus	0.37596346812051773
26.	A0090276	Nasopharynx	0.3571404036493111
27.	A0110234	Rectum	0.3421418045900255
28.	A0076424	Jejunum	0.2797947138605036
29.	A0293203	Tunica Media	0.2775
30.	A0095412	Oropharynx	0.2739805278599372
31.	A0055230	Epiglottis	0.2524645419034091
32.	A0026459	Arterioles	0.2450990009362834
33.	A0114520	Schwann Cells	0.24490246403081944
34.	A0100208	Pharynx	0.24017157023456717

35.	A0111706	Retinal Vein	0.23951562499999995
36.	A0091526	Neural Pathways	0.23925000000000002
37.	A0133305	Vocal Cords	0.2366532753086757
38.	A0031710	Blood–Brain Barrier	0.23536970160321882
39.	A0097632	Paranasal Sinuses	0.22836777000704916
40.	A0132130	Venules	0.2257269305734535
41.	A0030679	Biliary Tract	0.22514589082792213
42.	A0071331	Hypopharynx	0.21592129436644086
43.	A0020502	Afferent Pathways	0.21375000000000002
44.	A0089223	Myelin Sheath	0.21375000000000002
45.	A0072046	Ileocecal Valve	0.21375000000000002
46.	A0041253	Common Bile Duct	0.21261816605349632
47.	A0120521	Submandibular Gland	0.21183750000000007
48.	A0030639	Bile Ducts	0.20691838842975213
49.	A0075111	Intestinal Mucosa	0.20651574763651181
50.	A0361663	Microglia	0.20004375000000005
51.	A0046842	Dentition	0.1975353520653529
52.	A0035256	Cardiovascular System	0.1974462318831202
53.	A0126021	Thyroid Cartilage	0.19548828125000003
54.	A0090238	Nasal Mucosa	0.18782157131752758
55.	A1361065	Goblet Cells	0.18642857142857144
56.	A0032830	Bronchi	0.1861601545801232
57.	A0097950	Parotid Gland	0.1825749079245671
58.	A0111716	Retinal Vessels	0.18187500000000006
59.	A0362462	Superior Cervical Ganglion	0.18187500000000004
60.	A1378497	Pericytes	0.181875
61.	A0075149	Intestines	0.1781881796805679
62.	A2782503	Ampulla of Vater	0.17575757575757578
63.	A0111426	Respiratory System	0.1737135232823781
64.	A0043688	Cricoid Cartilage	0.17125
65.	A0025809	Appendix	0.16816042346938778
66.	A1645360	Enterocytes	0.16438928571428574
67.	A0029322	Basilar Artery	0.16416666666666668
68.	A8398635	Anal Canal	0.16275000000000003
69.	A0035102	Cardia	0.16275000000000003
70.	A2793862	Upper Gastrointestinal Tract	0.1627403787878788
71.	A0044706	Cystic Duct	0.16159090909090912
72.	A0062141	Gastric Mucosa	0.16060148787202272
73.	A0102478	Pleura	0.15798241678099495
74.	A0036148	Cecum	0.15528214285714287
75.	A0048766	Digestive System	0.15000000000000002
76.	A0031733	Blood–Retinal Barrier	0.15000000000000002
77.	A0031735	Blood–Testis Barrier	0.15000000000000002
78.	A0091317	Nerve Net	0.15000000000000002
79.	A0091325	Nerve Tissue	0.15000000000000002
80.	A0091697	Neuroglia	0.15000000000000002
81.	A2788821	Lower Gastrointestinal Tract	0.15000000000000002
82.	A0075910	Islets of Langerhans	0.15000000000000002
83.	A0097105	Pancreatic Ducts	0.15000000000000002
84.	A0063663	Glottis	0.15000000000000002
85.	A0078247	Laryngeal Mucosa	0.15000000000000002
86.	A0078248	Laryngeal Muscles	0.15000000000000002
87.	A0057409	Extravascular Lung Water	0.15000000000000002
88.	A0107733	Pulmonary Alveoli	0.15000000000000002

89.	A0090247	Nasal Septum	0.15000000000000002
90.	A0055076	Ependyma	0.15000000000000002
91.	A0053018	Efferent Pathways	0.15000000000000002
92.	A0083898	Medial Forebrain Bundle	0.15000000000000002
93.	A1639005	Internal Capsule	0.15000000000000002
94.	A1045364	Perforant Pathway	0.15000000000000002
95.	A0094455	Oligodendroglia	0.15000000000000002
96.	A0028268	Axillary Artery	0.15000000000000002
97.	A0032384	Brachial Artery	0.15000000000000002
98.	A0293190	Tunica Intima	0.15000000000000002
99.	A0111660	Retinal Artery	0.15000000000000002
100.	A0030627	Bile Canaliculi	0.15000000000000002
101.	A8413866	Insulin−Secreting Cells	0.15000000000000002
102.	A0026813	Arytenoid Cartilage	0.15000000000000002
103.	A0094431	Olfactory Mucosa	0.15000000000000002
104.	A0818087	Vomeronasal Organ	0.15000000000000002
105.	A0056576	Ethmoid Sinus	0.15000000000000002
106.	A0061026	Frontal Sinus	0.15000000000000002
107.	A0118449	Sphenoid Sinus	0.15000000000000002
108.	A0126812	Tonsil	0.15000000000000002
109.	A0062894	Geniculate Ganglion	0.15000000000000002
110.	A0092668	Nodose Ganglion	0.15000000000000002
111.	A0118621	Spiral Ganglion	0.15000000000000002
112.	A0128510	Trigeminal Ganglion	0.15000000000000002
113.	A0027876	Auditory Pathways	0.15000000000000002
114.	A0094438	Olfactory Pathways	0.15000000000000002
115.	A0133010	Visual Pathways	0.15000000000000002
116.	A0362714	Visceral Afferents	0.15000000000000002
117.	A0108203	Pyramidal Tracts	0.15000000000000002
118.	A0116694	Sinus of Valsalva	0.15000000000000002
119.	A1040957	Paneth Cells	0.15000000000000002
120.	A0056110	Esophagogastric Junction	0.15000000000000002
121.	A0062127	Gastric Fundus	0.15000000000000002
122.	A0108127	Pyloric Antrum	0.15000000000000002
123.	A0120513	Sublingual Gland	0.15000000000000002
124.	A0123286	Taste Buds	0.15000000000000002
125.	A0363583	Olfactory Receptor Neurons	0.15000000000000002
126.	A0119444	Stellate Ganglion	0.15000000000000002
127.	A2782458	Sphincter of Oddi	0.15000000000000002
128.	A0054770	Enterochromaffin Cells	0.15000000000000002

Appendix D

Results of *K*-Means - Ranked

GenCluster 1

```
(  0,   7) ---> Blood-Testis Barrier
(  1,   9) ---> Gastrointestinal Tract
(  2,  10) ---> Liver
(  3,  24) ---> Retinal Vessels
(  4,  25) ---> Bile Ducts
(  5,  31) ---> Islets of Langerhans
(  6,  76) ---> Tongue
(  7,  77) ---> Bile Canaliculi
(  8, 102) ---> Sinus of Valsalva
(  9, 103) ---> Common Bile Duct
( 10, 125) ---> Appendix
```

GenCluster 2

```
(  0,   0) ---> Cardiovascular System
(  1,   1) ---> Digestive System
(  2,   8) ---> Biliary Tract
(  3,  12) ---> Larynx
(  4,  13) ---> Lung
(  5,  14) ---> Nose
(  6,  15) ---> Pharynx
(  7,  17) ---> Ganglia
(  8,  22) ---> Arteries
(  9,  26) ---> Gallbladder
( 10,  28) ---> Upper Gastrointestinal Tract
( 11,  29) ---> Lower Gastrointestinal Tract
( 12,  33) ---> Glottis
( 13,  34) ---> Laryngeal Mucosa
( 14,  35) ---> Laryngeal Muscles
( 15,  38) ---> Pulmonary Alveoli
( 16,  39) ---> Nasal Cavity
( 17,  40) ---> Nasal Mucosa
( 18,  41) ---> Nasal Septum
( 19,  43) ---> Hypopharynx
( 20,  44) ---> Nasopharynx
( 21,  47) ---> Afferent Pathways
```

(22, 50) ——> Internal Capsule
(23, 57) ——> Arterioles
(24, 60) ——> Brachial Artery
(25, 64) ——> Venules
(26, 66) ——> Retinal Artery
(27, 69) ——> Duodenum
(28, 70) ——> Esophagus
(29, 73) ——> Jejunum
(30, 74) ——> Dentition
(31, 84) ——> Goblet Cells
(32, 85) ——> Olfactory Mucosa
(33, 86) ——> Vomeronasal Organ
(34, 87) ——> Ethmoid Sinus
(35, 88) ——> Frontal Sinus
(36, 90) ——> Sphenoid Sinus
(37, 91) ——> Tonsil
(38, 95) ——> Trigeminal Ganglion
(39, 96) ——> Auditory Pathways
(40, 98) ——> Visual Pathways
(41, 99) ——> Visceral Afferents
(42, 109) ——> Rectum
(43, 110) ——> Anal Canal
(44, 112) ——> Esophagogastric Junction
(45, 113) ——> Cardia
(46, 114) ——> Gastric Fundus
(47, 115) ——> Gastric Mucosa
(48, 117) ——> Ileocecal Valve
(49, 118) ——> Parotid Gland
(50, 121) ——> Taste Buds
(51, 126) ——> Sphincter of Oddi
(52, 127) ——> Enterochromaffin Cells

GenCluster 3

(0, 2) ——> Respiratory System
(1, 3) ——> Nervous System
(2, 4) ——> Blood Vessels
(3, 5) ——> Blood–Brain Barrier
(4, 18) ——> Nerve Net
(5, 19) ——> Nerve Tissue
(6, 20) ——> Neural Pathways
(7, 21) ——> Neuroglia
(8, 23) ——> Microcirculation
(9, 30) ——> Mouth
(10, 36) ——> Bronchi
(11, 37) ——> Extravascular Lung Water
(12, 45) ——> Oropharynx
(13, 46) ——> Ependyma
(14, 48) ——> Efferent Pathways
(15, 51) ——> Perforant Pathway
(16, 52) ——> Astrocytes
(17, 53) ——> Oligodendroglia
(18, 54) ——> Schwann Cells
(19, 55) ——> Microglia

(20, 56) ——> Aorta
(21, 59) ——> Basilar Artery
(22, 61) ——> Tunica Intima
(23, 62) ——> Pericytes
(24, 63) ——> Capillaries
(25, 65) ——> Tunica Media
(26, 89) ——> Maxillary Sinus
(27, 92) ——> Geniculate Ganglion
(28, 93) ——> Nodose Ganglion
(29, 97) ——> Olfactory Pathways
(30, 100) ——> Pyramidal Tracts
(31, 101) ——> Myelin Sheath

GenCluster 4

(0, 6) ——> Blood–Retinal Barrier
(1, 27) ——> Intestines
(2, 58) ——> Axillary Artery
(3, 67) ——> Retinal Vein
(4, 68) ——> Intestinal Mucosa
(5, 71) ——> Stomach
(6, 72) ——> Ileum
(7, 83) ——> Thyroid Cartilage
(8, 104) ——> Cystic Duct
(9, 105) ——> Paneth Cells
(10, 106) ——> Enterocytes
(11, 107) ——> Cecum
(12, 108) ——> Colon
(13, 111) ——> Ampulla of Vater
(14, 116) ——> Pyloric Antrum
(15, 119) ——> Sublingual Gland
(16, 124) ——> Superior Cervical Ganglion

GenCluster 5

(0, 11) ——> Pancreas
(1, 32) ——> Pancreatic Ducts
(2, 42) ——> Paranasal Sinuses
(3, 75) ——> Salivary Glands
(4, 78) ——> Insulin–Secreting Cells
(5, 79) ——> Vocal Cords
(6, 80) ——> Arytenoid Cartilage
(7, 81) ——> Cricoid Cartilage
(8, 82) ——> Epiglottis
(9, 120) ——> Submandibular Gland

GenCluster 6

(0, 16) ——> Pleura
(1, 49) ——> Medial Forebrain Bundle
(2, 94) ——> Spiral Ganglion
(3, 122) ——> Olfactory Receptor Neurons
(4, 123) ——> Stellate Ganglion

Appendix E

Results of PoBOC - Ranked

GenCluster 1

```
( 0,   9) ——> Biliary Tract
( 1,  10) ——> Gastrointestinal Tract
( 2,  11) ——> Liver
( 3,  12) ——> Pancreas
( 4,  13) ——> Larynx
( 5,  14) ——> Lung
( 6,  15) ——> Nose
( 7,  17) ——> Pleura
( 8,  30) ——> Lower Gastrointestinal Tract
( 9,  39) ——> Pulmonary Alveoli
(10,  41) ——> Nasal Mucosa
(11,  44) ——> Hypopharynx
(12,  45) ——> Nasopharynx
(13,  76) ——> Salivary Glands
(14,  85) ——> Goblet Cells
(15,  90) ——> Maxillary Sinus
(16, 109) ——> Colon
(17, 121) ——> Submandibular Gland
```

GenCluster 2

```
( 0,   1) ——> Cardiovascular System
( 1,   2) ——> Digestive System
( 2,   3) ——> Respiratory System
( 3,   4) ——> Nervous System
```

GenCluster 3

```
( 0,  23) ——> Arteries
( 1,  37) ——> Bronchi
( 2,  52) ——> Perforant Pathway
( 3,  54) ——> Oligodendroglia
( 4,  57) ——> Aorta
( 5,  63) ——> Pericytes
( 6,  64) ——> Capillaries
( 7,  69) ——> Intestinal Mucosa
```

(8, 70) ——> Duodenum
(9, 71) ——> Esophagus
(10, 72) ——> Stomach

GenCluster 4

(0, 5) ——> Blood Vessels
(1, 46) ——> Oropharynx
(2, 93) ——> Geniculate Ganglion

GenCluster 5

(0, 40) ——> Nasal Cavity
(1, 41) ——> Nasal Mucosa
(2, 85) ——> Goblet Cells

GenCluster 6

(0, 53) ——> Astrocytes
(1, 55) ——> Schwann Cells
(2, 56) ——> Microglia

GenCluster 7

(0, 12) ——> Pancreas
(1, 39) ——> Pulmonary Alveoli
(2, 75) ——> Dentition
(3, 118) ——> Ileocecal Valve
(4, 120) ——> Sublingual Gland
(5, 121) ——> Submandibular Gland

GenCluster 8

(0, 8) ——> Blood–Testis Barrier
(1, 25) ——> Retinal Vessels
(2, 26) ——> Bile Ducts
(3, 32) ——> Islets of Langerhans
(4, 73) ——> Ileum
(5, 77) ——> Tongue
(6, 103) ——> Sinus of Valsalva
(7, 104) ——> Common Bile Duct
(8, 126) ——> Appendix

GenCluster 9

(0, 24) ——> Microcirculation
(1, 66) ——> Tunica Media
(2, 67) ——> Retinal Artery

GenCluster 10

(0, 59) ——> Axillary Artery
(1, 90) ——> Maxillary Sinus

GenCluster 11

(0, 42) ——> Nasal Septum
(1, 88) ——> Ethmoid Sinus
(2, 90) ——> Maxillary Sinus

GenCluster 12

(0, 27) ——> Gallbladder
(1, 105) ——> Cystic Duct
(2, 106) ——> Paneth Cells
(3, 113) ——> Esophagogastric Junction

GenCluster 13

(0, 74) ——> Jejunum
(1, 89) ——> Frontal Sinus

GenCluster 14

(0, 43) ——> Paranasal Sinuses
(1, 80) ——> Vocal Cords
(2, 81) ——> Arytenoid Cartilage
(3, 82) ——> Cricoid Cartilage
(4, 83) ——> Epiglottis

GenCluster 15

(0, 68) ——> Retinal Vein
(1, 84) ——> Thyroid Cartilage
(2, 115) ——> Gastric Fundus

GenCluster 16

(0, 6) ——> Blood–Brain Barrier
(1, 62) ——> Tunica Intima

GenCluster 17

(0, 21) ——> Neural Pathways
(1, 22) ——> Neuroglia
(2, 101) ——> Pyramidal Tracts

GenCluster 18

(0, 33) ——> Pancreatic Ducts
(1, 79) ——> Insulin–Secreting Cells
(2, 80) ——> Vocal Cords

GenCluster 19

(0, 105) ——> Cystic Duct
(1, 107) ——> Enterocytes
(2, 108) ——> Cecum

(3, 125) ——> Superior Cervical Ganglion

GenCluster 20

(0, 28) ——> Intestines
(1, 111) ——> Anal Canal
(2, 113) ——> Esophagogastric Junction
(3, 114) ——> Cardia

GenCluster 21

(0, 20) ——> Nerve Tissue
(1, 49) ——> Efferent Pathways

GenCluster 22

(0, 95) ——> Spiral Ganglion
(1, 124) ——> Stellate Ganglion