FAKULTÄT
FÜR !NFORMATIK

Faculty of Informatics

Concept Mapping or Indexing of (Biomedical) Text

BACHELORARBEIT

im Rahmen des Studiums

Medizinische Informatik

eingereicht von

Edita Rados

Matrikelnummer 0725643

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung
Betreuer/in: Mag. Dr.rer.soc.oec. Katharina Kaiser

Wien, 25.04.2012

Contents

Abstract

Getting to accurate biomedical information is becoming more difficult as the amount of available data constantly increases. These data are usually stored and organized in different ways, as different languages or different terminologies are used, etc. Natural Language Processing (NLP) is a research area which main purpose is to help automated systems to understand the semantic of human language and enable easier usage of data. With the help of NLP, different techniques are used to identify specific noun phrases and to map them to their corresponding concepts. In this work we will take a closer look on some of the most important concept mapping methods in the biomedical area.

1. Introduction

Obtaining key information has always been a long and arduous task. The mere process of searching through information required a lot of time and patience, further made difficult by scarcity of information outside of closed circles. A special problem was ‘useless’ information that made up parts of result of surveys. This was a huge problem in the field of medicine, where time and correct information play an important role.

In order to improve search results, there had to be a way to centralize information, “separate” and “filter” it. That is how terms “concept mapping” and “concept indexing” came to be. Concept mapping maps (biomedical) text to concepts from vocabularies. This is how searches are limited only to information related to certain notions. The techniques based on concept mapping can be used for data mining and information retrieval of any domain with adequate knowledge sources and not only specific for the biomedical domain.

Indexing of biomedical literature was once done by hand and by human indexers. Human indexing consists of reviewing the complete text of an article and assigning descriptors that represent the central concepts as well as other topics that are discussed to a significant extent [1]. Together with constant growth of information, automatic indexing became a necessity.

Different methods for mapping and indexing have been developed, taking accuracy, acceptable time, and memory usage in account. Some of these methods are MetaMap, MGREP, and Medical Text Indexer (MTI).

2. Natural Language Processing (NLP)

Since the advent of computers, scientists have had one desire: to make computers understand their needs. Natural Language Processing (NLP) is a research area whose main purpose is to make automated systems (computers) understand natural language. A natural language is a language that is spoken, written by humans for general-purpose communication [2]. A language can be seen as a system that is composed out of symbols and rules (or grammar) [2]. The Symbols are combined to convey new information and the rules govern the manipulation of symbols.

Developing a system that completely understands natural language is very difficult because words can have more than one meaning and to fully understand their meaning we need to know the context of the text. Also the same concept could be expressed with different words. These ambiguous words are mapped to the defined terms with the help of the context of the sentence.

Natural Language Processing (NLP) techniques, such as part-of speech taggers and parsers, have been applied to capture biological entities and complex relationships among them from text [3]. Using NLP techniques to extract coded data from free text allows the use of natural language as the input medium.

NLP is being used in many areas, including medicine. The use of NLP in medical area is very important because it enables a new level of functionality for health care applications [4]. NLP systems are being used for different applications, such as decision support, surveillance of infectious diseases, research studies, automated encoding, quality assurance, indexing patient records, etc. [4]. Use of NLP systems reduces mistakes and “exceptions”, which could happen during the revision, comparison of the documents or searching for new information [4].

Another significant advantage is that NLP technology can be used to standardize reports from diverse institutions and applications because the same automated system will uniformly encode clinical information that occurs in heterogeneous reports, thereby facilitating interoperability [4].

The Figure 2.1 shows components of a generic clinical application that utilizes NLP extraction technology [4]

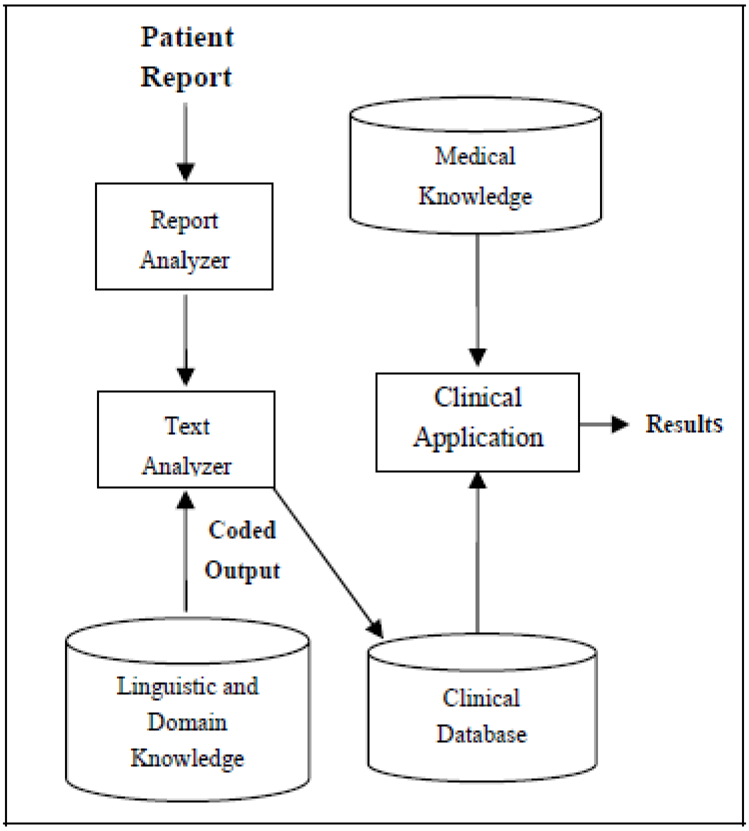


Figure 2.1 Components of a generic NLP Application in the Clinical Domain [4]

The patient report first goes through Report Analyzer, which identifies segments and handles textual irregularities (i.e. tables, domain-specific abbreviations). After that, the Text Analyzer is used for information extraction. This is the core component of the NLP engine. The Text Analyzer uses linguistic knowledge associated with syntactic and semantic features, and a conceptual model of the domain to structure and encode the clinical information and to generate output, which is then stored for subsequent use, generally in a structured coded clinical database [4]. In the end the automated clinical application can use the structured data and if needed it can also use additional medical knowledge.

3. Biomedical Terminologies/Ontologies

An ontology is a very important subject in the biomedical world [5]. Generally, it is referred to as explicit specification of conceptualization [6].

Ontological terminologies are frequently described as enabling resources in text mining systems [7]. These resources are used to support tasks such as entity recognition (i.e., the identification of biomedical entities in text), and relation extraction (i.e., the identification of relationships among biomedical entities) [7].

Biomedical ontologies describe the concepts and relationships that can exist and formalize the terminology of a domain [8]. On the other side the purpose of biomedical terminologies are to collect the names of entities employed in the biomedical domain [9].

Biomedical terminologies have become very useful tools for information retrieval.

They improve text-driven access by supplying a standard vocabulary for indexing information in a specific domain. [10]. Biomedical terminologies and ontologies are also welcome as external resources supporting text mining tasks, such as entity recognition or relation extraction [10].

We will take closer look to some of the biomedical terminology systems such as UMLS and GALEN that are being used in the biomedical area.

3.1 UMLS

The Unified Medical Language System (UMLS)[1] was developed by National Library of Medicine (NLM)[2] in order to provide a bridge between concepts stored in different vocabularies. Because different names can describe the same concept it was necessary to develop a standard which allows the effective retrieval of information. The UMLS was developed almost 20 years ago and today it contains over 900 000 concepts from all types of vocabularies and over 12 million relations between these concepts [11].

[1] http://www.nlm.nih.gov/research/umls/

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

The UMLS system contains three subsystems: the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon (Figure 3.1).

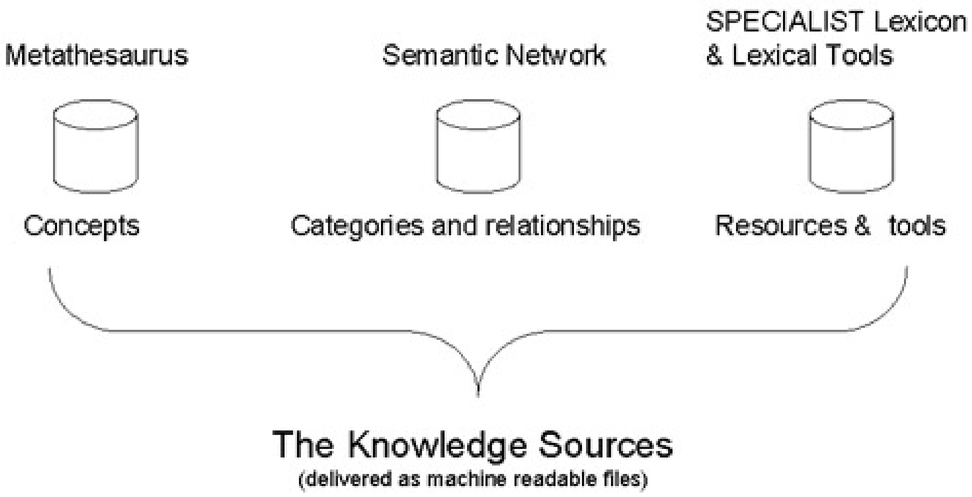

Metathesaurus
Semantic Network
SPECIALIST Lexicon & Lexical Tools
Concepts
Categories and relationships
Resources & tools
The Knowledge Sources
(delivered as machine readable files)

Figure 3.1 UMLS Subsystems [12]

UMLS concepts are not only interrelated but may also be linked to external resources such as GenBank [11].

With UMLS we can easily access biomedical literature, medical record, facts, etc. In the UMLS, synonymous terms are clustered together to form a concept and concepts are linked to other concepts by means of various types of relationships. Inter-concept relationships are either inherited from the structure of the source vocabularies or generated specifically by the editors of Metathesaurus.

3.1.1 Metathesaurus

The Metathesaurus is a large, multi-purpose, and multi-lingual vocabulary database made by connecting dozens of medical vocabularies, called thesauri, such as Medical Subject Headings (MeSH is used to index MEDLINE)[3], SNOMED CT

[3] http://www.ncbi.nlm.nih.gov/mesh

(Clinical Coding System)[4], and so on [12].

The Metathesaurus contains all kinds of biomedical concepts, health data, classifications and definition from controlled vocabularies. It is used in patient care, health services billing, public health statistics, indexing and cataloging biomedical literature, and/or basic, clinical, and health services research [12].

The most important parts of the Metathesaurus are concepts. The Metathesaurus assigns several types of unique, permanent identifiers to the concepts and concept names it contains, in addition to retaining all identifiers that are present in the source vocabularies. The Metathesaurus concept structure includes concept names, their identifiers, and key characteristics of these concept names [13].

The Metathesaurus has four levels of structure:

1. Concept Unique Identifiers (CUI)
2. Lexical (term) Unique Identifiers (LUI)
3. String Unique Identifiers (SUI)
4. Atom Unique Identifiers (AUI)

Each concept has its own concept unique identifier (CUI) and each concept name is pointing to all lexical terms with same meaning, and each unique name from each language has its unique String identifier (SUI). As the basic building blocks of the Metathesaurus, concept names or strings are called “atoms”. Every occurrence of a string in each source vocabulary is assigned a unique atom identifier (AUI). If exactly the same string appears twice in the same vocabulary a unique AUI is assigned for each occurrence. When the same string appears in multiple source vocabularies, it will have AUIs for every time it appears as a concept name in each of those sources. All of these AUIs will be linked to a single string identifier (SUI), since they represent occurrences of the same string [13]. LUI is used to link strings that are lexical variants. Lexical variants are detected using the Lexical Variant Generator (LVG) program, one of the UMLS lexical tools [12]. Lexical tools are usually a set of Java programs that allow the user to manage lexical variation in biomedical text.

[4] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

In Figure 3.2 we can see how the Methatesaurus can serve as a link between not only vocabularies, but also the subdomains it represents.

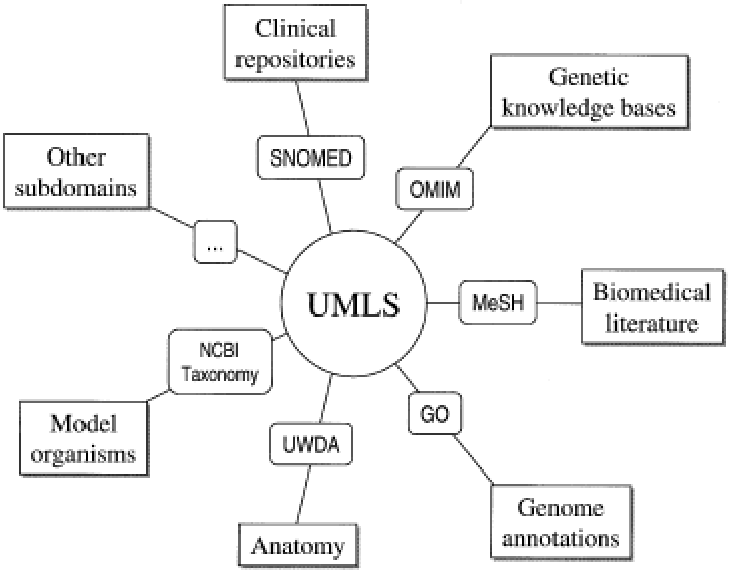


Figure 3.2 Various subdomains integrated in the UMLS [11]

3.1.2 Semantic Networks

The Semantic Network[5] is composed of “semantic types” which create a consistent categorization of concepts in the UMLS Metathesaurus, and a set of important relationships or semantic relations, which exist between these semantic types. The Semantic Network contains 135 Semantic Types and 54 relationships. There are two types of relationships: hierarchical and non-hierarchical relationships.

Hierarchical relationships are:

- isa
- part of

[5] http://semanticnetwork.nlm.nih.gov/

Non-hierarchical relationships are:

- physically related to
- spatially related to
- temporally related to
- functionally related to
- conceptually related to

The primary link between the semantic types is the 'isa' link and it is used for deciding on the most specific semantic type available for assignment to a Metathesaurus concept [12].

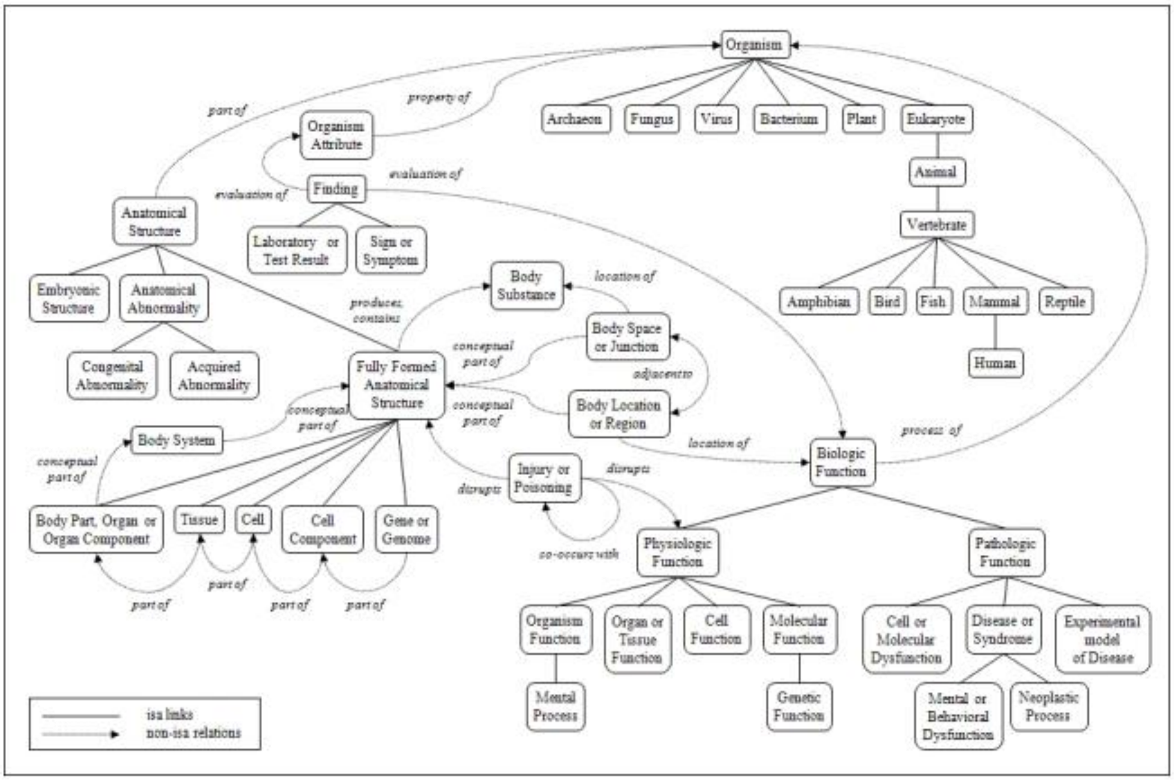


Figure 3.3 *Portion of the UMLS Semantic Network showing semantic types linked by relationships and the hierarchical structure of the semantic types [14].*

This Figure 3.3 shows the portion of a Semantic Network with its relations (hierarchical and associative), which exist between the semantic types.

The Semantic Network is a useful resource to categories medical entities and to acquire relationships among them [3].

3.1.3 SPECIALIST lexicon

The SPECIALIST Lexicon[6] provides the system with needed lexical information. It contains the most biomedical terms in English. The lexicon entry for each word or term records the syntactic, morphological, and orthographic information needed by the SPECIALIST NLP System [15].

The lexicon has entries for each spelling or it's spelling variants in a particular part of speech. These items can be "multi-words" terms if the item should be a lexical item and can be determined by its presence as a term in English or medical dictionaries in MeSH (medical thesauri).

The unit of a lexical record consists of slots and fillers which are in a frame structure.

The records have

- a "base=" slot for the base form,
- an optional slot for "spelling_variants=" and
- an "entry=" slot for a unique identifier (EUI)
- a cat= slot indicating part of speech

The records are delimited by curly braces ({…})

Example [16]:

{base=anesthetic spelling_variant=anaesthetic entry=E0354094 cat=noun variants=reg variants=uncount} {base=anesthetic spelling_variant=anaesthetic entry=E0330019 cat=adj variants=inv position=attrib(3) position=pred stative}

The SPECIALIST Lexicon is available as an open source resource as part of the SPECIALIST NLP tools [15]

[6] http://www.nlm.nih.gov/pubs/factsheets/umlslex.html

3.1.4 Controlled Vocabularies and information resources

The Metathesaurus drains its information from more than 100 controlled vocabularies.

A controlled vocabulary (CV) can be defined as "*organized lists of words and phrases, or notation systems, that are used to initially tag content, and then to find it through navigation or search*" [17].

On the next pages some of the most important controlled vocabularies of the Metathesaurus are described, such as MeSH, ICD-10, SNOMED CT, and RxNorm [18] [19].

3.1.4.1 MeSH

MeSH[7] is the biggest and most used medical thesaurus which was developed by the National library of Medicine (NLM)[8]. MeSH is NLM's controlled vocabulary for subject indexing and searching of journal articles in MEDLINE[9], and books, journal titles, and non-print materials in NLM's catalog [20].

Medline is the largest online data base that contains almost all published articles in the field of biomedicine. These articles can be accessed on the web for free with the help of PubMed [21].

MeSH terms are arranged in a hierarchical categorized manner called MeSH Tree Structures and are updated annually. The MeSH main parts are [22]:

- Subject headings (also known as Main headings, or descriptors - are used to describe what an article or book is "about." That is, as index terms they provide an indication of the major topics under consideration [20].
- Subheadings (also known as Qualifiers) - are used to narrow the specific focus of a main topical heading to a particular aspect of the subject [23].
- Supplementary Concept Records (also known as Substance Names)

[7] http://www.ncbi.nlm.nih.gov/mesh

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

[9] http://www.nlm.nih.gov/bsd/pmresources.html

The descriptors are tagged with letters, and every entry is granted with one letter. Each of these letters stands for a specific category.

There are 16 different categories in which the descriptors are grouped:

- Anatomic Terms [A]
- Organisms [B]
- Diseases [C]
- Drugs and Chemicals [D]
- Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
- Psychiatry and Psychology [F]
- Biological Sciences [G]
- Physical Sciences [H]
- Anthropology, Education, Sociology and Social Phenomena [I]
- Technology and Food and Beverages [J]
- Humanities [K],
- Information Science [L]
- Persons [M]
- Health Care [N]
- Publication Characteristics [V]
- Geographic Locations [Z]

Each of these categories is further divided in subcategories and every entry gets its own number which describes hierarchical classification, for example [24]:

C- Disease
C8 Respiratory Tract Diseases
C8.127 Bronchial Diseases
C8.127.108 Asthma

Subheadings are used to describe the specific aspects of the MeSH heading that are pertinent to the article. Most MeSH terms, especially those from Categories A, B, C and D, will usually be indexed with one or more subheadings. Usually no more than three subheadings are needed; however the indexer can use more than three, if the article demands it [25]. Currently, MeSH contains 83 different subheadings. There are three different forms in which the subheadings can appear: full name, two-letter abbreviation, or indexing abbreviation [25].

Supplementary Concept Records are edited and added to MeSH daily, and the majority of these records are related to chemicals and drugs [20]. Each Supplementary Concept Record is mapped to one or more MeSH records and the indexing of Supplementary Concept Record terms is identical to the indexing of MeSH descriptors.

3.1.4.2 SNOMED CT

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)[10] was developed over the past years at the College of American Pathologists. SNOMED CT has its roots in SNOP (Systematized Nomenclature of Pathology) [26] that was developed in 1965. It was followed by SNOMED I, SNOMED II, SNOMED 3, and SNOMED RT (Reference Terminology). SNOMED CT was formed in 1999 by the convergence of SNOMED RT and the United Kingdom's Clinical Terms Version 3 (formerly known as the Read Codes) [26]. In 2007 SNOMED CT was acquired by IHTSDO (International Health Terminology Standards Development Organization).
SNOMED CT provides “a common language” for use by clinical specialists for clinical notes [27]. It provides an extensible foundation for expressing clinical data in local systems, for interoperability, and for use in data warehouses. It is designed for clinical documentation and reporting [29].

SNOMED CT is a concept-based terminology. With the help of SNOMED CT concepts can be related to each other, grouped, and analyzed according to different criteria. Numeric codes (the SNOMED CT identifier) identify every instance of the three core building blocks: concepts, descriptions, and relationships. Each concept represents a single specific meaning; each description associates a single term with a concept; and each relationship represents a logical relationship between two concepts [29].

[10] http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html

SNOMED CT content is organized into nineteen independent hierarchies:

- Clinical finding
- Physical force
- Procedure
- Event
- Observable entity
- Environments/geographical locations
- Body structure
- Social context
- Organism
- Situation with explicit context
- Substance
- Staging and scales
- Pharmaceutical/biologic product
- Linkage concept
- Specimen
- Qualifier value
- Special concept
- Record artifact
- Physical object

SNOMED provides a rich set of inter-relationships between concepts. Hierarchical relationships define specific concepts as children of more general concepts. For example, “kidney disease” is defined as a kind of “disorder of the urinary system.” In this way, hierarchical relationships provide links to related information about the concept. This example shows that kidney disease has a relationship to the concept that represents the part of the body affected by the disorder [30].

For diseases/disorders, SNOMED CT uses relationships between concepts to provide logical, computer readable definitions of medical concepts. There are several types of relationships described or modelled in SNOMED CT [30]:

1) "Is A" Relationship: Creates a hierarchical relationship between concepts, relating specific concepts to a more general category.
2) "Finding Site" Relationship: identifies the part of the body affected by the specific disorder or finding.
3) "Causative agent" Relationship: Identifies the direct cause of the disorder or finding. The causative agent is the bacterium, virus, toxin or environmental agent that causes the disorder.
4) "Associated morphology" Relationship: I nbzdentifies the abnormal physical condition that is characteristic of a given disorder or finding.

For medical and surgical procedures, SNOMED CT applies following relationships [30]:

1. "Is A" Relationship
2. "Procedure Site" Relationship: The "Procedure Site" relationship identifies the part of the body acted on by the procedure.
3. "Method" Relationship: identifies the kind of procedure that is being carried out.

5) "Direct Morphology" Relationship: Identifies the abnormal physical condition that is being directly addressed by the procedure.
6) "Direct Device" Relationship: Identifies the device that is involved in the core operation of the procedure.
7) "Using" Relationship: Refers to any instrument, equipment or energy that is used to perform the procedure.

SNOMED CT terminology has been integrated into UMLS Metathesaurus. With this step the Metathesaurus gained one very important vocabulary which, linked with other Metathesaurus vocabularies, brought better use and understanding of clinical data. SNOMED CT has been converted to the common UMLS format, its concept names have been connected to those already in the Metathesaurus, and its content has been assigned UMLS identifiers, semantic types, etc. [31].

SNOMED CT is being used in more than 30 countries all over the world and its use will continue to grow as it represents an important point in the biomedical development. The clinical domain is becoming increasingly important as clinical information systems try to implement problem-oriented documentation and maintain a longitudinal or historical record of patients’ problems. It has been found out that SNOMED CT has around 80% to 90% coverage of problems included in medical problem list domains of individual enterprise reference terminologies [32].

3.1.4.3 ICD-10

The International Classification of Diseases and Related Health Problems (ICD)[11] developed by the WHO (World Health Organization) is a classification system used in medical documentation. Every medical entry (e.g., diagnosis of the diseases) should be entered according to the standards provided by ICD-10.

Just like SNOMED CT, ICD-10 also has a mono-hierarchical structure. There are 22 main categories in which all medical diseases and states are divided. These categories are further divided into subcategories until the defined concrete diagnose is reached.

ICD-10 is being used worldwide and it was translated in many languages but only English and German can be used in UMLS. Efforts are being made to enable the complete use of ICD-10 in the UMLS. The problem of using ICD-10 in UMLS is that some information is lacking because many data items from ICD-10 are missing [33]. For instance, the inclusion terms or terms from the Alphabetical Index are only available when they are referred to from other vocabularies and they are not linked to ICD-10. Users of the UMLS will get the impression that many clinical terms are missing from ICD-10. Manifestation and codes are not distinguished from etiology codes. Furthermore, all information on dual coding is entirely missing [33]. Dual Coding means adding both ICD-10 and ICD-9 codes simultaneously to the record [34]. As a result, healthcare organizations can avoid unnecessary missed reimbursement opportunities and strengthen staff proficiency with ICD-10 [35].

3.1.4.4 RxNorm

RxNorm[12] was developed by National Library of medicine (NLM), it is a standardized nomenclature for clinical drugs. RxNorm ensures that all clinical drugs and drug

[11] http://www.dimdi.de/static/de/klassi/diagnosen/icd10/

[12] http://www.nlm.nih.gov/research/umls/rxnorm/

delivery devices have the same standard names. This is very important when there is an exchange of data between facilities or even inside the same facilities.

As RxNorm is part of the UMLS Metathesaurus, the clinical drug names used in RxNorm are connected to other drug names which are stored in other Metathesaurus vocabularies. Each drug name contains some or all of the following “parts”: ingredients, strengths, and/or form, so that there should not be any confusion because every combination of these “parts” represents a unique RxNorm name.

Within RxNorm, generic and branded normalized forms are related to each other and to the names of their individual components by a well-defined set of named relationships. For example "Acetaminophen 500 MG Oral Tablet" is related to "Acetaminophen 500 MG Oral Tablet [Tylenol]," and both have relationships to "Acetaminophen, Acetaminophen 500 MG," and "Oral Tablet." Within the UMLS Metathesaurus, "Acetaminophen 500 MG Oral Tablet" and "Acetaminophen 500 MG Oral Tablet [Tylenol]" will each be linked to different names that are used for these entities in other vocabularies [36].

3.2. GALEN

Generalized Architecture for Languages, Encyclopaedias and Nomenclatures in medicine (GALEN)[13] is a European Union project that is used to provide open source clinical system.

At early stages of the GALEN Programme, the researchers developed GRAIL [37] the concept modelling language which was used to test some clinical demonstrator projects. GRAIL is defined as GALEN Representation and Integration Language.
In the later stages of GALEN, more tools and techniques were developed. One of them was GALEN Common Reference Model which is used for GALEN terminology server.

[13] http://www.opengalen.org/

Open GALEN has been set up as a not-for-profit Dutch Foundation by the universities of Manchester and Nijmegen [7] to provide access to the GALEN Common Reference Model and to the description of GALEN technology. The current version of Open GALEN (Dec. 2002) [7] contains about 25.000 concepts, and it essentially provides building blocks required to describe the concepts.

In Figure 3 Open Galen levels have been presented. At top level we have “Phenomenon” and “Modifier Concept”. The Modifier concept [7] is used to distinguish concepts that represent things with independent existence (physical objects, for example) from dependent concepts such as modifiers (Mild severity), states (Pathological state) or roles (Infective role).

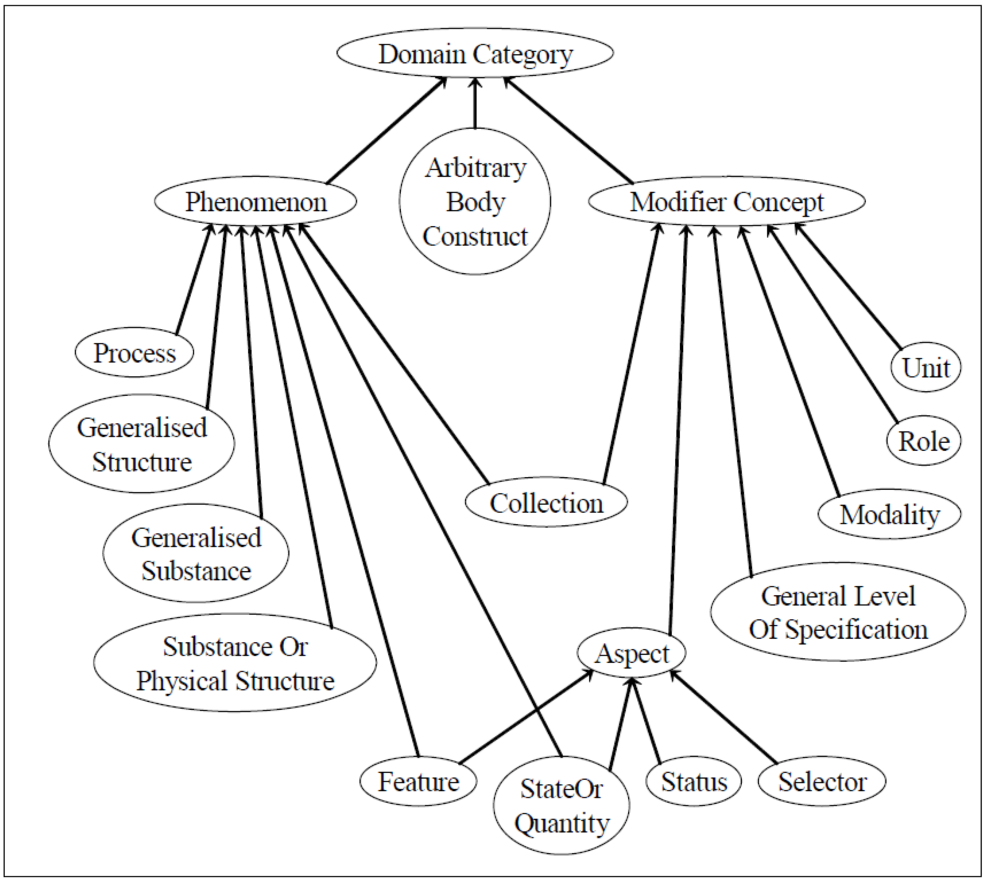


Figure 3.4 *Top levels in Open Galen [7]*

Galen differs from other biomedical systems because it tries to solve existing problems with different approach. In GALEN, the main concepts can be described with the help of simple forms and it has the ability to compose many descriptions from manageable number of base concepts [38]. One other difference is that GALEN uses Common reference model instead of standard coding system.

UMLS can be seen as ontology whereas GALEN is more a terminology. GALEN originated the idea of a terminology server and is participating actively in the CorbaMed (division of CORBA devoted to healthcare) effort at standardizing the software interface [38].

4. Mapping and Indexing tools

The availability of large medical online collections, such as MEDLINE, brings new challenges to information and knowledge management. The Researchers need services that are able to process the metadata of diverse resources to annotate and index them with concepts from appropriate ontologies, and that can enable the researchers to locate resources related to particular ontology concepts [50].

In the past years, multiple systems for automatic annotation of large data resources have been developed. The main goal of these systems is the identification of concepts based on the text analysis of documents [69]. On the following pages, several existing systems used for mapping and indexing of biomedical data will be reviewed.

4.1 MetaMap

MetaMap[14] was developed by the Lister Hill National Center for Biomedical Communications at the institute for Natural Language Programming. Because of the constant growth of information, there was a need for indexing and mapping of this information for easier use.

The main purpose of MetaMap is to map biomedical text to Metathesaurus concepts. MetaMap is using a NLP- based approach and computational linguistic techniques. It can be used for data mining and information retrieval of any domain with adequate knowledge sources and not only specific for the biomedical domain.

In the early phases of MetaMap the mapping was done manually in order to get accurate information on how the automatic MetaMap mapping algorithm should work. 310 phrases were manually mapped to Metathesaurus in 1992 [39].

[14] http://metamap.nlm.nih.gov/

Depending on how well the phrase was mapped, they were classified into one of four categories:

1. Simple match: This means that the phrase entered is the same as the string in the Metathesaurus.
2. Complex match: The whole word does not exactly match the string of the Metathesaurus, but each part of noun has its own simple match in the Metathesaurus.
3. Partial match: The noun phrase maps to a Metathesaurus string in such a way that at least one word, of either the noun phrase or the Metathesaurus string (or both) does not participate in the mapping. Partial matches have the following variations:
    a. Normal partial match: For example, *liquid crystal thermography* maps to Thermography where the mapping does not involve *liquid crystal* [39].
    b. Gapped partial match: The words *ambulatory monitoring* will be mapped to AMBULATORY CARDIAC MONITORING. The gap is “Cardiac”, because it does not occur in the first sentence [39].
    c. Overmatch: The noun entered can have one or more matches in the Metathesaurus. For example, “Application” has many overmatches like Job Application, Heat/Cold Application and Medical Informatics Application [39].
4. No match: There is not a single part of the noun that matches a string in the Metathesaurus.

Just like the manual mapping, there are also a few steps which need to be done in order to obtain an automatical mapping as shown in the Figure 4.1. Each step will be explained separately further down in the text.

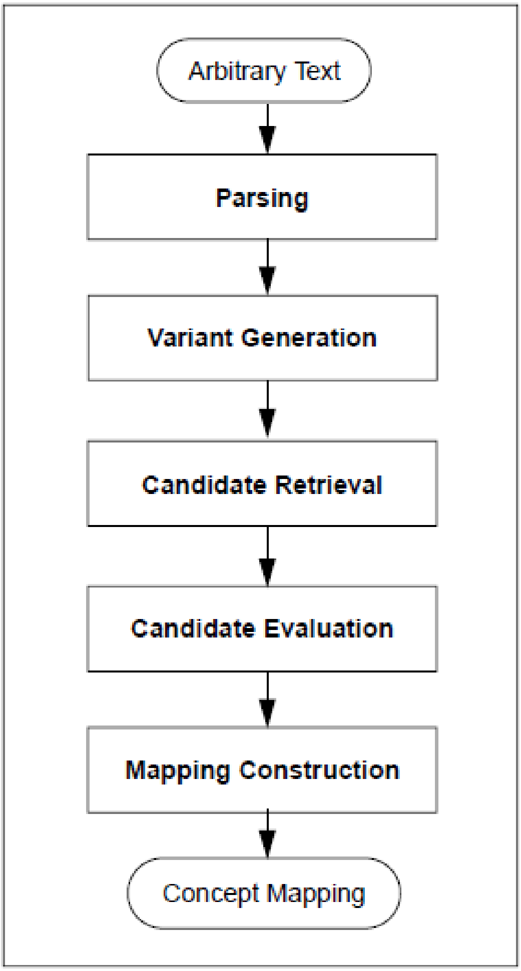


Figure 4.1 *MetaMap processing [40]*

Parsing

The parsing is done by the Specialist parser which segments biomedical text into simple noun phrases. The parser produces a high-level syntactic analysis. The parser uses the Xerox part-of-speech tagger which assigns syntactic tags (e.g., noun, verb) to words not having a unique tag in the SPECIALIST lexicon [41]. The incorrect tags are minimal and do not have much influence on the subsequent steps so they can be ignored.

Example: if we have ocular complications of myasthenia gravis as a text fragment, the parser will detect two noun phrases. One is “ocular compilations” and the other is “of myasthenia gravis”.

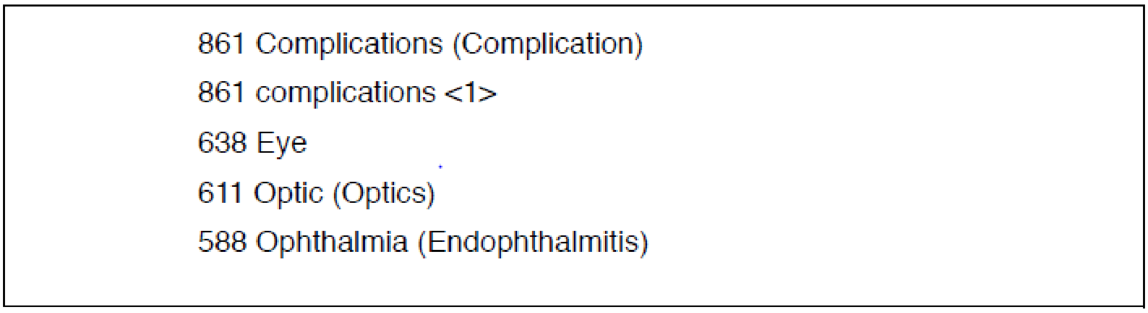
Variant Generation

After the parser has detected the noun phrases, for each phrase a variant is being generated. A Variant consists of one or more noun phrase words (called generator) together with all of its spelling variants, abbreviations, acronyms, synonyms, inflectional and derivational variants, and meaningful combinations of these [40]. A variant generator is any meaningful subsequence of words in the phrase where a subsequence is meaningful if it is either a single word or occurs in the SPECIALIST lexicon [40].

For example, if we have a phrase ”of obstructive sleep apnea” then “of” is filtered out and variant generators would be the multi-word items “obstructive sleep apnea”, “sleep apnea”, and the single words “obstructive”, “sleep” and “apnea” [42].

Candidate Retrieval

The Metathesaurus candidates are sets for noun phrases which consist of a set of strings which are containing at least one variant computed for input phrase. The Metathesaurus candidates for the phrase “ocular complications” are shown in Figure 4.2.



861 Complications (Complication)
861 complications <1>
638 Eye
611 Optic (Optics)
588 Ophthalmia (Endophthalmitis)

Figure 4.2 *Metathesaurus Candidates for ocular complications [39]*

The candidates are ordered with an “evaluation function”, which will be described later. In this example the best candidates are “Complication” and “complications<1>”. The other candidates are the variants of the word “ocular” and are described with similarities of this word.

Candidate Evaluation

In this phase each candidate is evaluated against the input text by computing a mapping from the phrase word to the candidate word. The the mapping strength is calculated using a “linguistically principled evaluation” function for all of these candidates by consistence of a weighted average of the following metrics:

1. Centrality (Involvement of Head): The centrality value is simply 1 if the string involves the head of the phrase and 0 otherwise. For the noun phrase ocular complications, Complications has centrality value 1; and Eye has value 0 [39].
2. Variation (Avg. of inverse distance score): The variation value estimates how much the variants in the Metathesaurus string differ from the corresponding words in the phrase. It is computed by first determining the variation distance for each variant in the Metathesaurus string. This distance is the sum of the distance values for each step taken during variant generation [39].
3. Coverage: The coverage value indicates how much of the Metathesaurus string and the phrase are involved in the match. In order to compute the value, the number of words participating in the match is computed for both the Metathesaurus string and the phrase [39].
4. Cohesiveness: The cohesiveness value emphasizes the importance of connected components. A connected component is a maximal sequence of contiguous words participating in the match. The connected components for both the Metathesaurus string and the phrase are computed [39].
5. Involvement: The involvement value is a replacement of the coverage value when word order is ignored. The strict word order implied by the matchmap is no longer followed. The involvement value for the phrase is the proportion of phrase words which can map to a Metathesaurus word whether or not they do according to the matchmap [39].

Mapping Construction

The final mapping consists of the combinations of the Metathesaurus candidates, which are the disjoined parts of the noun phrase. The highest scoring completed mapping represents the best interpretation of the original phrase in MetaMap. The phrase “ocular complications” consists of the concept “Ocular” and also of the concept “Complication” or the concept “Complications” [42]. This illustrates the problems of MetaMap which is ambiguity, because both concept have “complications” in one of their strings and therefore cannot be distinguished by the MetaMap.

There is also a Java implementation of MetaMap called MetaMap Transfer (MMTx).

MMTx is an effort to make the MetaMap program available to biomedical researchers in a generic, configurable environment [43]. Experiments comparing MetaMap and MMTx showed that there were some differences between MetaMap and MMTx [43]. It was found that MMTx does not handle parenthetical expressions correctly. MMTx attaches parenthetical expressions to the phrase left of them, rather than splitting them into a separate phrase [43]. MMTx and MetaMap returned candidates of the same ranking in different order. It was also found that MetaMap did not recognize some concepts whereas MMTx did [43]. The reason for that was that MetaMap includes a stop phrase feature that does not process phrases known to produce no results.

As from 2011, the use of MMTx has been suspended and all focus has been placed on further development of MetaMap [71].

4.2 AMTEx

AMTEx [44] is a medical document indexing method, specifically designed for the automatic indexing of documents in large medical collections, such as MEDLINE. AMTEx uses C/NC- value method, a hybrid linguistic/statistical term extraction method for extraction of multi-word and nested terms [66]. In this method, the text is first tokenized and tagged by a part-of-speech tagger. Subsequently, a set of rules and linguistic filters is used to identify in text candidate term phrases [44].

There are three linguistic filters:

- N+N
- (A | N)+N
- ((A | N) + | ((A | N)*(N P)?)(A | N)*)N

Where N is a noun, A is an adjective and P stands for a preposition.

Depending on which filters are being used the precision and recall of the system changes. When a closed filter is used, such as the first one, the precission of the system will be increased but the recall will be decreased, whereas an open filter, such as the last one will increase recall and decrease precision [44].

The AMTEX method has six stages [44]:

1. Multi-word Term Extraction: For the term extraction, AMTEx parses the document text, using the C/NC-value part-of-speech tagger and linguistic filters
2. Term Ranking: Extracted candidate terms are evaluated, first by C-value and subsequently by NC-value score. The final candidate term list is ranked by decreasing term likelihood. Top ranked terms are more important than terms ranked lower in the list and are more likely to be included in the final list of extracted terms.
3. Term Mapping: Candidate terms are mapped to terms of the MeSH Thesaurus (by applying simple string matching) so that after that the list of terms contains only MeSH terms.
4. Single-word Term Extraction: The C/NC-value tends to produce compound (multi-word) terms (it does not produce single-word terms). Often such terms include shorter terms (mostly single-word terms) which are also MeSH terms. Single-word terms are also extracted and are added to the candidate term list (the text is scanned and each word is checked against the MeSH vocabulary).
5. Term Variants: Term variants are included in the candidate term list. The C/NC-value implementation in AMTEx includes inflectional variants of the extracted terms. Also, MeSH itself can be used for locating variant terms, based on the MeSH term, Entry Terms property. However, only the stemmed term-forms are used in AMTEx since the full list of Entry Terms may contain terms that often are not synonymous.
6. Term Expansion: The list of terms is augmented with semantically (conceptually) similar terms from MeSH.

4.3 MetaCoDe

Another tool being used to map biomedical data to the UMLS Metathesaurus is MetaCoDe[15] IT was developed for the mapping of big biomedical collections that were written primarily in French but it is also used for text sources in English. This

[15] http://www.semantic-valley.org/en/

tool is still in development and it is an open source tool.

Currently there are two versions of MetaCoDe: MetaCoDe V0.1- a text mapping tool using UMLS entries, and MetaCoDe V0.2 also known as “MetaCoDe In GATE” which is a text mapping plugin for GATE (General architecture for text engineering)[16]. GATE is an open source platform based on Java used to solve text processing problems [47].

In comparison with other mapping tools MetaCoDe needs less time to compute the task and it uses less memory.

4.3.1 MetaCoDe V0.1

MetaCoDe v0.1 was developed to extract UMLS tags from biomedical text written in other languages beside English [45]. Till now it has been used for French text but it could easily be extended to other languages because of its simple logic and short programs (written in C++ and Pearl).

MetaCoDe involves the chaining of a set of distinct operations, as illustrated in Figure 4.3. The first bunch of operations is to prepare both the corpus and the necessary terminological resources (preliminary tasks); the last operation is the mapping itself [46].

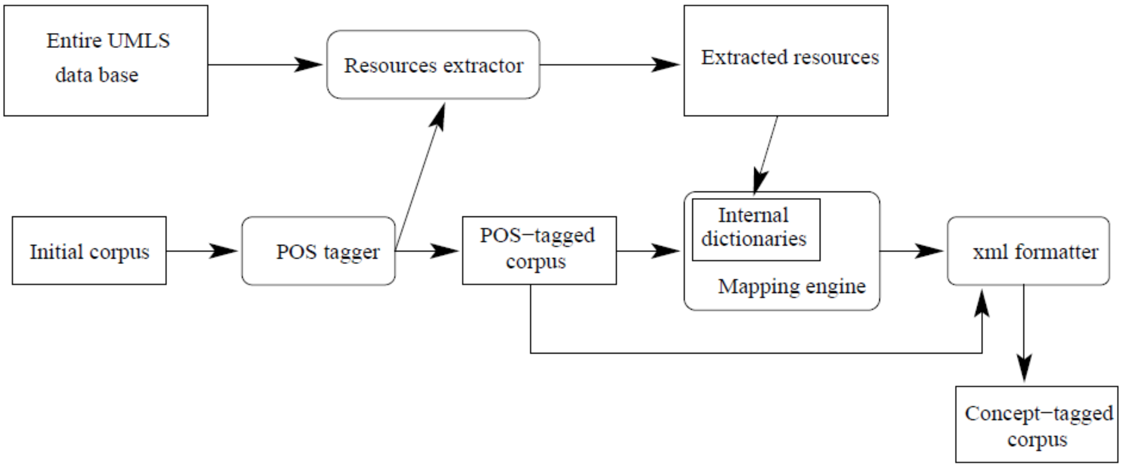


Figure 4.3 *MetaCoDe overall process [46]*

[16] http://gate.ac.uk

MetaCoDe v0.1 is still in its developing stage and because of lack of feedback it is progressing very slowly. There is no GUI, nor unique script, and a fair amount of preprocessing needs to be done to run the tagger [47].

4.3.2 MetaCoDe V0.2

Like its previous version MetaCoDe v0.2 is written in C++ and is used as dynamic library in GATE.

The main motivations for this implementation are:

- To operate at a high speed
- To create an open platform which can be integrated easily into other platforms
- To get a scalable and easily maintainable system
- To operate “on the fly” to access stream data from the Internet
- To work with limited computing resources
- To save resources for other processes

Along with tokenization and part-of-speech tagging, noun phrase chunking is achieved thanks to a transducer defined with a JAPE grammar [47]. JAPE is a Java Annotation Patterns Engine which allows us to recognize regular expressions in annotations on documents.

MetaCoDe Mapping Engine Algorithm has four steps [48]:

- Step 1: Noun Phrase Extraction (GATE PR). Based on POS extraction and JAPE grammar => each noun phrase is later represented as a bag of words $\{w_1, \ldots, W_n\}$.
- Step 2: For each noun phrase, candidates SUI’s are selected based on the bag of words => $C_1 = \{sui_1, \ldots, sui_p\}$.
- Step 3: a lattice on C1 is built; the lattice is pruned to keep the SUIs with wider coverage of the noun phrase without being too specific => C_2
- Last Step: for each element of C_2, its corresponding CUI and semantic types from the UMLS database are retrieved.

The version v0.2 has also two graphical components:

- The MetaCoDe annotations Sets viewer
- The MetaCoDe annotations viewer

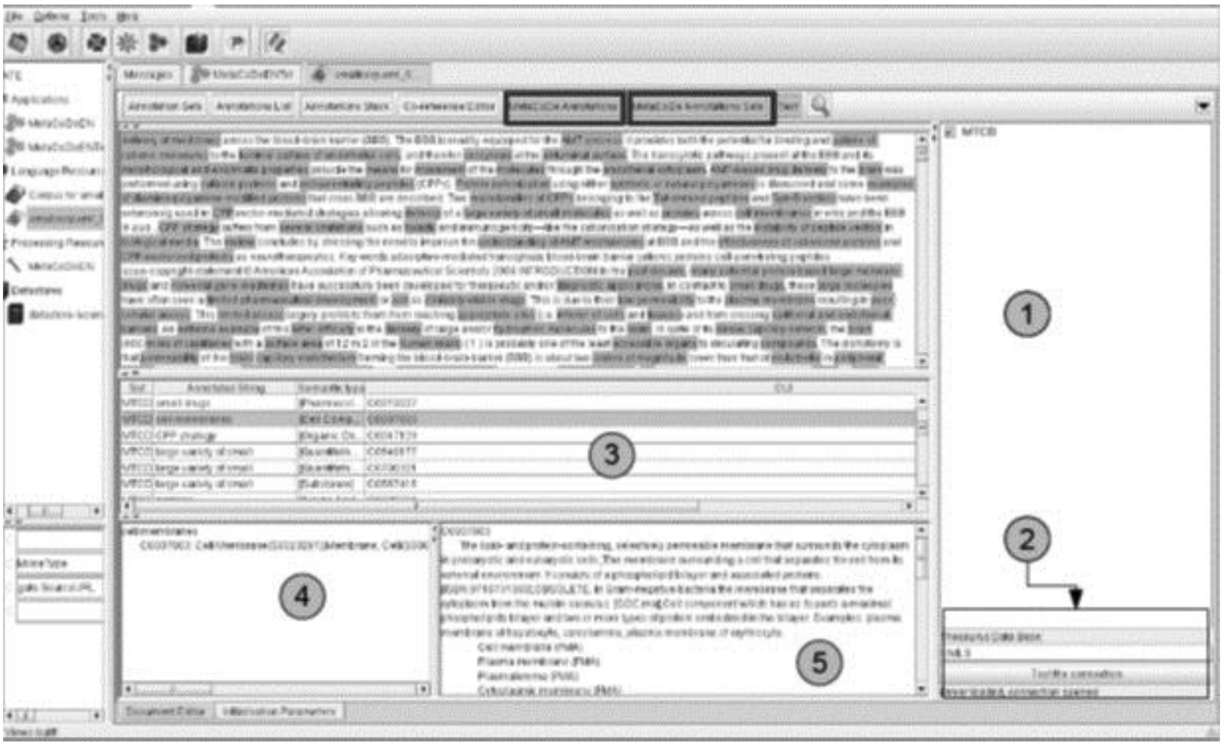


__Figure 4.5__ MetaCoDe V0.2 GUI: 1-The MetaCoDe Annotations Sets Viewer; 2-Part of the MetaCoDe Annotations Sets Viewer, sets up the connection to an UMLS data base; 3-Part of the MetaCoDe Annotations Viewer, lists the annotations; 4-Part of the MetaCoDe Annotations Viewer, displays the concepts tagging a selected text fragment; 5-Part of the MetaCoDe Annotations Viewer, displays further [49].

The MetaCoDe Annotations Viewer is able to give the user more specific information about the tags. In particular it can use connections to the UMLS data base tables to fetch informations about semantic types, synonymous, and definitions of the extracted concepts [49]. Figure 4.5 shows a screenshot of the MetaCoDe plugin in GATE [49].

The core tagging process of the MetaCoDe algorithm is fast enough to process big corpora or to add extra algorithms for specific applications [48]. Because the core algorithm of the tagger is not depending on the application language, MetaCoDe can easily be used for other languages beside French. The only thing that needs to be done is to change the POS tagging rules and rewrite the JAPE grammar.

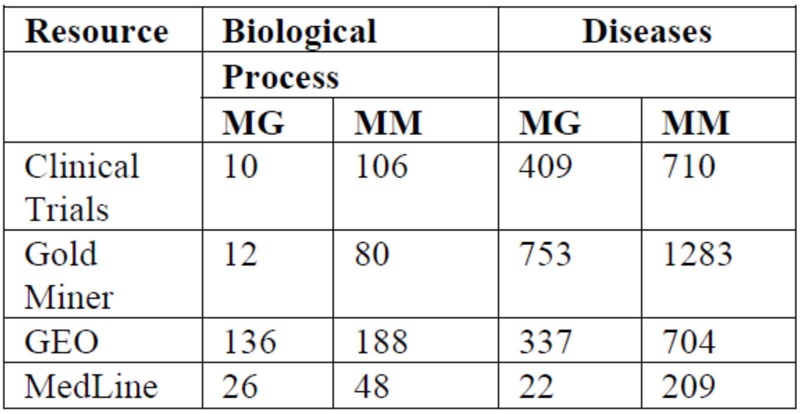
4.5 MGREP

MGREP [70] is a mapping tool developed by the National Center for Integrative Biomedical Informatics (NCIBI) at the University of Michigan. It is used for concept recognition with a high degree of customizability vis-à-vis dictionaries and resources.

MGREP algorithm is a string-matching algorithm which means that it finds text segments inside a string based on already existing pattern.

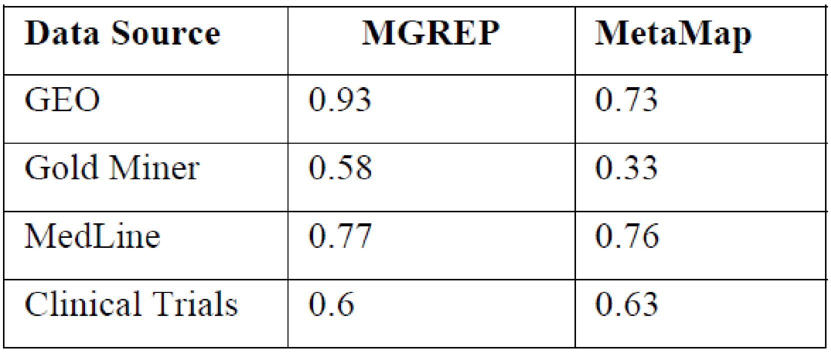
A study, done at National Center of Biomedical Ontology [50] at Stanford University has shown that MGREP in comparison to MetaMap, recognizes a lower number of unique

Concepts (see Figure 4.6) but In general, MGREP has a higher precision in recognizing Biological Processes (see Figure 4.7 and Figure 4.8). However, MGREP identifies a large number of concepts that are redundant – concepts recognized at the same position in the input string – and overall the number of unique concepts recognized is less than with MetaMap [50].



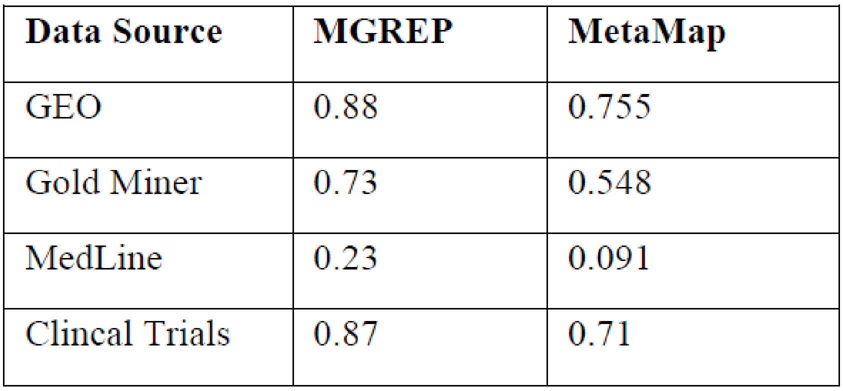
Resource	Biological Process		Diseases	
	MG	MM	MG	MM
Clinical Trials	10	106	409	710
Gold Miner	12	80	753	1283
GEO	136	188	337	704
MedLine	26	48	22	209

Figure 4.6 *Total number of concepts recognized by MGREP and MetaMap across all resources using the biological process and diseases dictionaries. MG=MGREP; MM =MetaMap [50]*



Data Source	MGREP	MetaMap
GEO	0.93	0.73
Gold Miner	0.58	0.33
MedLine	0.77	0.76
Clinical Trials	0.6	0.63

Figure 4.7. *Precision of MGREP and MetaMap using Biological Processes as the dictionary [50].*



Data Source	MGREP	MetaMap
GEO	0.88	0.755
Gold Miner	0.73	0.548
MedLine	0.23	0.091
Clincal Trials	0.87	0.71

Figure 4.8. *Precision of MGREP and MetaMap using the 'diseases' dictionary which contains UMLS concepts that are of semantic type 'disease or syndrome' [50].*

Because of its speed and precision MGREP has been used to develop first online based annotation tool called "Open Biomedical Annotator" which will be explained further in this document.

4.6 Open Biomedical Annotator (OBA)

Nowadays, almost all biomedical data can be found on the internet. This enables faster data exchange, but because of the large number of data, getting the required information has become more difficult. Another problem that appears with this sort of data is that the majority of the information is not written in accordance to existing annotation standards that would enable easier and faster data research. Reasons that these annotations are not being done is that the user himself feels overwhelmed (overthrown) with an enormous number of ontologies which often change and also because all these annotations have to be done manually that acquires a lot of time.

The National Canter for Biomedical Ontology has developed a web-based tool named “Open Biomedical Annotator” (OBA)[17]. This tool allows automatic annotation of biomedical text by tagging data (textual metadata) to biomedical ontology concepts.

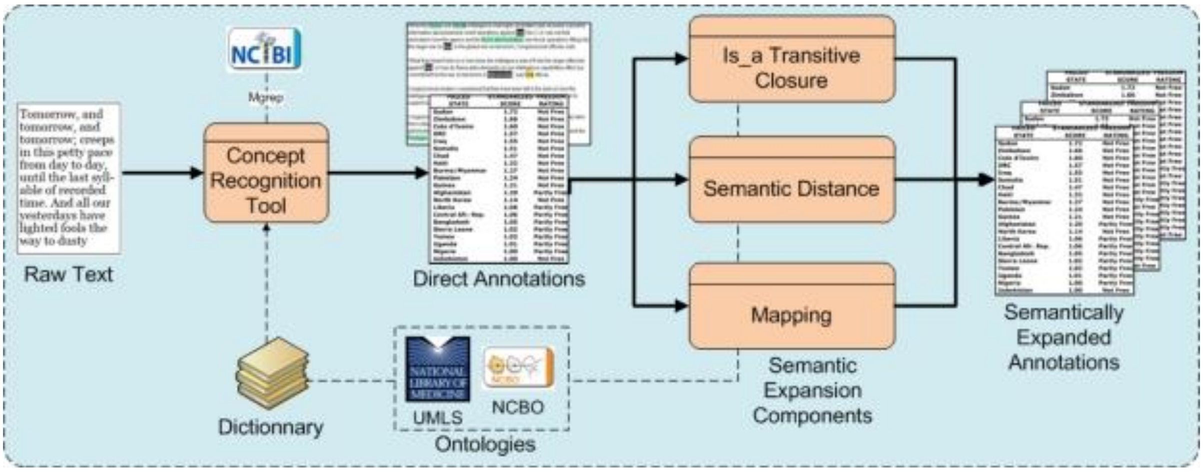

NCIBI
Mgrep
Tomorrow, and tomorrow, and tomorrow, creeps in this petty pace from day to day, until the last syllable of recorded time. And all our yesterdays have lighted fools the way to dusty
Raw Text
Concept Recognition Tool
Direct Annotations
Is_a Transitive Closure
Semantic Distance
Mapping
Semantic Expansion Components
Semantically Expanded Annotations
Dictionnary
UMLS
NCBO
Ontologies

***Figure 4.9** OBA web service workflow [28]*

The Open Biomedical Annotation workflow consists of two steeps (see Figure 4.9):

1. Raw text submitted by the user is given as an input into the Concept Recognition Tool. As a concept recognizer OBA uses MGREP mapping tool. As another input this tool receives the Dictionary, which contains strings that identify the biomedical ontology concepts. The dictionary is constructed by accessing biomedical ontologies and pooling all concept names or other string forms, such as synonyms or labels that syntactically identify concepts. As a concept recognition tool result we get direct annotations. The tool recognizes concepts by using string matching on the dictionary [28].

[17] http://bioportal.bioontology.org/annotator

2. After recognizing the direct annotation the Semantic Expansion components are being used in order to expand the direct annotations. There are Semantic Expansion components [51]:
    - An is_a transitive closure component traverses an ontology parent-child hierarchy to create new annotations with parent concepts of the concepts involved in direct annotations. For instance, if data are directly annotated with the concept melanoma from NCI Thesaurus, this semantic expansion component can generate new annotations with concepts skin tumor and neoplasms because NCI Thesaurus provides the knowledge that melanoma is_a skin tumor and skin tumor is_a neoplasms. The maximum level in the hierarchy to use is parameterizable.
    - A semantic distance component uses a given notion of concept similarity to obtain related concepts and create new annotations. For instance, if a text is directly annotated with the concept melanoma from Mesh, this semantic expansion component can generate new annotations with concepts apudoma and neurilemmoma because Mesh specifies these three concepts as siblings in the hierarchy.
    - An ontology-mapping component creates new annotations based on existing mappings between different ontologies. For instance, if a text is directly annotated with the concept NCI/C0025202 (melanoma in NCI Thesaurus), this semantic expansion component can generate new annotations with concepts SNOMEDCT/C0025202 (melanoma in SNOMED-CT) and 38865/DOID: 1909 (melanoma Hunan disease) because the UMLS and the NCBO BioPortal provides the mapping information.

As a result we get a set of semantically expanded annotations. The system can classify these semantically expanded annotations by accounting for the frequency with which a concept has been identified directly by the concept recognizer or by semantic expansion components [52].

4.7 SAPHIRE

SAPHIRE (Semantic and Probabilistic Heuristic Information Retrieval Environment) [54] is concept based automated indexing system developed at the Division of Medical Informatics and Outcomes Research, of Oregon Health Sciences University.

SAPHIRE uses a non-syntactic pattern-matching approach to extract concepts represented in controlled vocabularies from free text [53]. The text entered can be any medical document or user query to a retrieval system. As an output SAPHIRE delivers a single exact matching Metathesaurus concept or a ranked list of partially matching concepts [53].

The SAPHIRE algorithm, after receiving the input starts by breaking the input string into individual words. To each word weight is assigned, based on their frequency in the document and infrequency in other documents. If they occur with a frequency above a specified cut-off in the Metathesaurus, they are designated as common. The purpose of designating words as common is to reduce the computational overload for words which are occasionally important in some terms but occur frequently in others [54]. For each word in the input string, a list of Metathesaurus terms in which the word occurs is constructed. Only those terms that occur in one or more of the non-common words in the input string will be added to the Metathesaurus term lists for common words. Once the term lists for each word are created, a master term list is created that contains any term which occurs in one or more individual word lists. Terms in which less than half of the words occur in the input string are discarded.

The terms are then weighted based on formula that gives weight to terms that are longest, have the highest proportion of words from the term in the string, and have the words of the term occurring in close proximity to each other. Terms that match all the words in the input string exactly are given additional weight [54].

The algorithm process is very fast, with query processing occurring nearly instantaneously for a few-word query and under 10-15 seconds for a 10-15 word sentence [68].

In the past, SAPHIRE has been used only for English terms in the Metathesaurus. However, since 1998 there has been an internationalization of the SAPHIRE concept matching system, allowing it to accept text input and provide Metathesaurus concept output in any of six languages: English, German, French, Russian, Spanish, and Portuguese [54].

SAPHIRE has been used to identify concepts in electronic medical records [54]. But its main use recently has been to provide access to index terms in the Clinic Web catalogue of clinically-oriented pages on the World Wide Web [54].

A key problem with SAPHIRE's algorithm is that it may return multiple matching concepts for a text segment which, due to partial matching, are incorrect concept mappings [55]. The disambiguation problem was handled by adding a semantic type filter, which filters out concepts belonging to a particular semantic type [55].

4.8 MTI

The Medical Text Indexer (MTI)[18], developed by NLM, is an automated text processing system that derives ranked lists of MeSH terms to describe the content of medical journal citations using knowledge from the Unified Medical Language System and from MEDLINE [56].

The top of the MTI diagram consists of three paths (see Figure 4.5):

- MetaMap indexing Algorithm – used for discovering UMLS concepts. The first step is to apply the MetaMap program to a body of text, like the Title or abstract fields of MEDLINE. After that, the discovered UMLS concepts are ranked according to the MMI ranking function. This function presents a product of a frequency factor and relevance factor. The relevance factor is a weighted average of four components: a MeSH tree depth factor, a word length factor, a character count factor, and a MetaMap score factor [57].
- PubMed Related Citations algorithm - indirectly computes a ranked list of MeSH headings for an arbitrary body of text. The neighbors of the text, related citations, are those citations in Medline that are the most similar to it. The

[18] http://ii.nlm.nih.gov/mti.shtml

terms recommended by this path are extracted from the MeSH fields of those citations [58].

- Trigram – is a phrase matching algorithm for identifying phrases that have a high probability of being synonyms

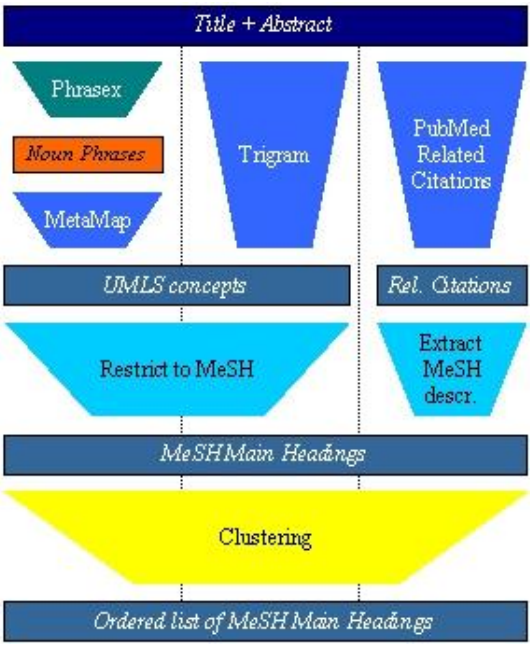


Figure 4.9 *Methods and processes used in MTI [60]*

The Restrict to MeSH algorithm is used to find the MeSH terms most closely related to each of the MetaMap identified UMLS concepts [59].

The last steep of MTI is to create a final ranked list of recommended indexing terms by weighting the ranked lists of MeSH headings produced by each of the indexing paths and combining them using the Clustering process.

4.9 IndexFinder

IndexFinder [61] is a concept mapping algorithm used for generating UMLS concepts in real time applications. The IndexFinder uses syntactic and semantic filtering to exclude the irrelevant concepts from the concepts list calculated by permuting the set of words in the input text.

Before the algorithm can be used on a certain text, the text needs to be preprocessed because the Indexfinder uses the UMLS normalized string table which only supports certain types of abbreviation [61]. In this process the terms in text are normalized, undefined and ambiguous abbreviations are detected and the stop terms are removed in order to increase the accuracy of the extraction. After that the remaining terms are mapped to their base form [61]. For each unique term, all UMLS concept phrases containing the term are retrieved. Each retrieved phrase maintains its length and the count of terms matched so far. After all terms have been evaluated, the retrieved phrases are then evaluated based on their counts. Concepts are extracted for indexing where concept phrases have all matching terms with the source phrase [55].

In order to focus on delivering the results that are most useful for the users, IndexFinder uses 6 different filters to improve results from which the first three are applied during the mapping process and other three are used for further pruning the candidate phrases.

The Mapping process filters are:

- Symbol Type filter: to specify the symbol types of interests.
- Term Length filter: to specify the length limitation of candidate phrases.
- Coverage filter: to specify the coverage condition for a candidate phrase.

The Pruning filters are:

- Subset filter: to remove phrases if they are subsets of some other phrases.
- Range filter: to remove a phrase if the phrase is found from words in the input text to exceed a specific distance.
- Semantic filter: to remove the phrases of semantic types that the user is not interested in.

An empirical analysis done by developers, showed that IndexFinder algorithm can process text at a throughput of 43 KB/second, which is much faster than NLP-based approaches like MetaMap [61].

Other experiments showed that while NLP-based approaches tend to break a sentence into small fragments, the IndexFinder considers all the possible combination of words in the input unit as long as valid in UMLS. That means that the results delivered by IndexFinder contain concepts which are more specific than those delivered by NLP-based approaches.

4.10 SENSE

SENSE (Search with New Semantics) [62] program is used to map user's quarries to relevant Medical Subject Headings (Mesh terms). This is done by transforming words and phrases in the users' queries into primary conceptual components (semantic factors) and comparing these components with those of the MeSH vocabulary. It can be said that the purpose of SENSE is to build a list of suggested MeSH terms [55].
The Figure 4.10 shows the process of mapping an input phrase to a MeSH term. There are three main parts of sense program: Medical Language Knowledge Base, a Semantic Analyzer, and a MeSH Semantic Index.

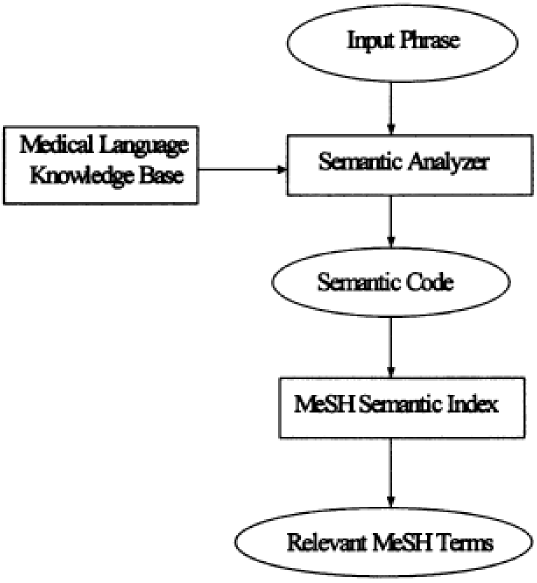


Figure 4.10 *Structure of the SENSE program [62]*

The Medical Language Knowledge Base is used to describe how concepts are expressed in medical language [62]. The Medical Language Knowledge Base together with The Semantic Analyzer replaces what the user types with its prime semantic factors. Prime semantic factor represents the concept which is so simple that there is no need to split it further. The semantic factors produced for a particular word or phrase is called "semantic code." The output of the SENSE program is a suggested list of MeSH terms which can be used to perform a search.

Next example shows input phrases followed by MeSH terms suggested by SENSE [62]:
CANCER ORIGINATING IN THE BRONCHIAL TREE
*carcinoma, bronchogenic
*bronchial neoplasms
*carcinoma
*carcinogens
*neoplasms
*bronchi
*wounds and injuries
*trees

The purpose of SENSE is to build a list of suggested MeSH terms. No effort is made to identify the best matching concept [55].

4.11 MedLEE

A Medical Language Extraction and Encoding System MedLEE [63] is used to extract, structure, and encode clinical information in textual patient reports and to translate the information to terms in a controlled vocabulary, such as the UMLS or SNOMED [63].

MedLEE structure consists of two parts: the programming components and knowledge sources (Figure 4.11). This structure allows MedLEE to be used in different clinical domains. This can be accomplished by creating new domain-specific knowledge sources while the programming components are left as they are [64].

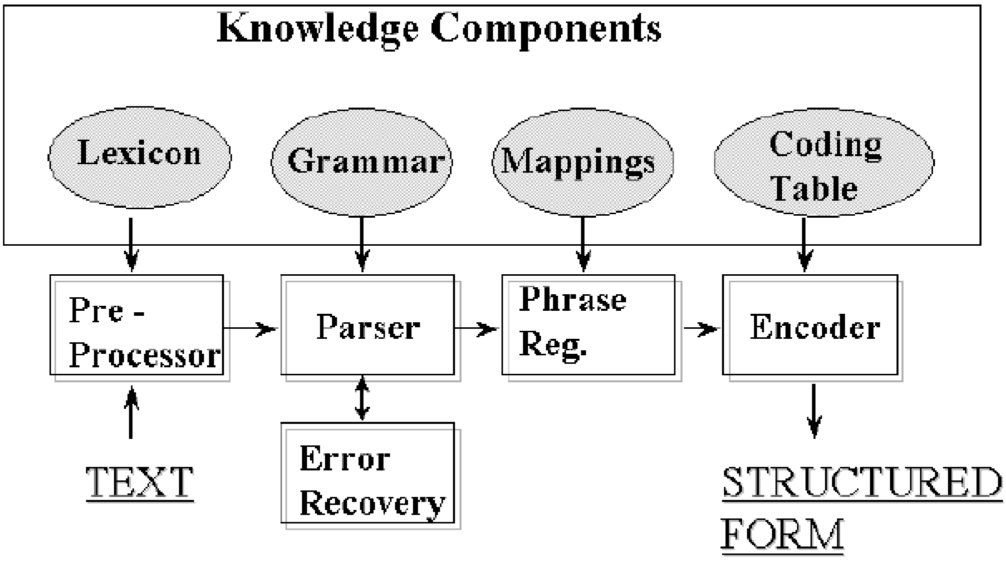


Figure 4.11 shows overview of the components used by MedLEE. The ovals represent knowledge components and the rectangles programming components [64].

The text was first processed by the pre-processor which defines the sections of the report, and identifies individual sentences. It does lexical lookup to identify and categorize single and multiword phrases in each sentence, and to determine the target output forms [64]. The output represents the list of elements and a separate list that contains the categories and target forms. The elements in the list of elements represent either a single word or a list of words that constitutes an atomic phrase [64].

After that the text is passed on to the parser which recognizes well-formed structures and generates target forms based on grammar rules and categories which are assigned to the words of the sentence. If a parse of the sentence is not obtained, but the sentence contains relevant clinical information, the error-recovery component is activated. This component uses various strategies to break up the sentence into fragments and to parse the segments [64]

Next step is to compose phrases that have been separated in the sentence. After that the encoder maps the target forms generated by the previous phase into a specified coded controlled vocabulary (e.g. UMLS) [64].

It allows for the automation and simplification of decisions when integrated into a clinical information system [63]. MedLEE has also been applied to improve safety through intervention. For example, MedLEE detects patients at high risk for having active tuberculosis and recommends respiratory isolation [63].

5. Conclusion

Concept mapping, particularly in the last decade, gained an important significance in the field of biomedical science. A large number of tools have been developed but only a few of them managed to achieve greater progress in providing interoperability between different biomedical terminologies. With increasing importance of the biomedical area further research in this area is needed.

The development of UMLS allowed connection of a large number of different vocabularies into UMLS Metathesaurus and contributed to more efficient sharing and reuse of biomedical data. The number of vocabularies incorporated into UMLS Metathesaurus is growing constantly and we can say that UMLS plays and will play an essential role in the processing of biomedical data.

The MetaMap mapping tool was developed for mapping biomedical text to the UMLS Metathesaurus but this tool has been improved so much that it can be used for data mining and information retrieval of any domain with adequate knowledge sources and not only for specific biomedical domain [67]. Unfortunately till now the tool has only been used in the biomedical domain.

Besides MetaMap, a large number of mapping and indexing tools have been developed. Mapping tools such as MetaCoDe and MGREP have fast response time with low memory usage. Studies have shown that MGREP can outperform the MetaMap when the precision is in question with a small exception being that MetaMap slightly outperformed MGREP for recognizing biological processes in http://ClinicalTrials.gov [65]. MGREP has also shown to be much faster in performing the indexing tasks. Open Biomedical Annotator (OBA) based on MGREP enables fast data exchange online, and is the first online based annotation tool.
One of the motivations for the development of the MetaCoDe v2.0 was to save resources for other processes and to make it possible to operate "on the fly" to access stream data from the internet.

Besides mapping, Indexing of biomedical data plays an important role in getting the most out of the biomedical text. One of the tools used is AMTEx which is designed for indexing and retrieval of documents in large medical collections, such as MEDLINE. Comparing experiments showed that AMTEx outperforms MMTx in both precision and recall [66].

It would be difficult to decide which tool is “the best” as they all have limitations and advantages, but the most used one is MetaMap.

.

References

[1] C. W. Gay, M. Kayaalp, and A. R. Aronson. Semi-automatic indexing of full text biomedical articles. *In: AMIA Annu Symp Proc*, pages 271–275, 2005.

[2] RC Chakraborty. Natural Language Processing: Artificial Intelligence. AI Course Lecture 41. Created at June 01, 2010.
http://www.myreaders.info/10_Natural_Language_Processing.pdf

[3] Maria Taboada, Maria Meizoso, Diego Martınez, and José J. Des. Using lexical, terminological and ontological resources for entity recognition tasks in the medical domain. Lecture Notes in Computer Science, Volume 4924, pages 21–31, 2008.

[4] Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh. *Medical Informatics*: Semantic text parsing for patient records. *Springer*, pages 423-448, July 28, 2005.

[5] Bodenreider O, Smith B, and Burgun A. The Ontology-Epistemology Divide: A Case Study in Medical Terminology. *In: Varzi AC, Vieu L, editors. Proceedings of the Third International Conference on Formal Ontology in Information Systems (FOIS 2004),* pages 185-195, 2004.

[6] Marek Obitko. Ontologies and Semantic Web: Specification of Conceptualization. Retrieved at Februar 10, 2012. http://www.obitko.com/tutorials/ontologies-semantic-web/specification-of-conceptualization.html

[7] Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh. *Medical Informatics*: Biomedical Ontologies. Springer, pages 212-236, July 28, 2005.

[8] Hsinchun Chen, Sherrilynne S. Fuller, Carol Friedman, and William Hersh. *Medical Informatics*: Knowledge Management and Data Mining in Biomedicine. Springer, pages 3-34, July 28, 2005.

[9] Oliver Bodenreider. Lexical, terminological and ontological resources for biological text mining. In: *Ananiadou S, McNaught J, editors. Text mining for biology and biomedicine: Artech House*, pages 43-66, 2006.

[10] McCray AT, Bodenreider O, Malley JD, and Browne AC. Evaluating UMLS strings for natural language processing. *Proc AMIA Symp*, pages 448-452, 2001.

[11] Bodenreider, Olivier. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, Volume 32. Oxford University Press, pages 267-270, January 01, 2004.

[12] Michael Bauer. The Unified Medical Language System, Created at: September 22, 2009. Retrieved at March 13, 2012.
http://www.csd.uwo.ca/courses/CS9626b/papers_files/The_Unified_Medical_Language_System.pdf

[13] National Library of Medicine. UMLS Reference Manual: Metathesaurus. Retrieved at: April 03, 2012. http://www.ncbi.nlm.nih.gov/books/NBK9684/

[14] National Library of Medicine. UMLS Reference Manual: Semantic Network. Retrieved at: April 03, 2012. http://www.ncbi.nlm.nih.gov/books/NBK9679/

[15] U.S. National Library of Medicine. Library Catalogs & Services: Fact Sheet: SPECIALIST Lexicon. National Institute of Health. Retrieved at: February 6, 2012.
http://www.nlm.nih.gov/pubs/factsheets/umlslex.html

[16] National Library of Medicine. UMLS Reference Manual: SPECIALIST Lexicon and Lexical Tools. Retrieved at: February 03, 2012.
http://www.ncbi.nlm.nih.gov/books/NBK9680/

[17] Fred Leise , Karl Fast, and Mike Steckel. Boxes and Arrows. What is A Controlled Vocabulary. Created at December 16, 2002.
http://boxesandarrows.com/view/what_is_a_controlled_vocabulary_

[18] U.S. National Library of Medicine. Library Catalogs & Services: Biomedical Research & Informatics: UMLS: Knowledge Sources: Metathesaurus. National Institute of Health. Retrieved at: February 6, 2012.
http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html

[19] DIMDI. Deutsches Institut fuer Medizinische Dokumentation und Information. Klassifikationen: UMLS. Retrieved at February 6, 2012.
http://www.dimdi.de/static/de/klassi/mesh_umls/umls/index.htm

[20] Stuart J. Nelson, W. Douglas Johnston, and Betsy L. Humphreys. Library Catalogs & Services: MeSH. U.S. National Library of Medicine: National Institute of Health. Retrieved at: February 6, 2012.

[21] Dolf Trieschnigg, Piotr Pezik, Vivian Lee, Franciska de Jong, Wessel Kraaij, and Dietrich Rebholz-Schuhmann. MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics (Oxford, England),* Vol. 25, No. 11, pages 1412-1418, June 1, 2009.

[22] Harald Trost. Taxonomie und Ontologe. Institut für Artificial Intelligence, Wien, pages 1-25, 2011/2012. http://www.meduniwien.ac.at/user/harald.trost/lv/termont-ch1.pdf

[23] U.S. National Library of Medicine. About the NLM: Cataloging: MeSH Training Course: Module 4. National Institute of Health. Retrieved at: April 2, 2012. http://www.nlm.nih.gov/tsd/cataloging/trainingcourses/mesh/mod4_010.html

[24] U.S. National Library of Medicine. Medical Subject Headings: MeSH Tree Structures: C8 - DISEASES-RESPIRATORY. National Institute of Health. Created at: 2008. Retrieved at: January 8, 2012. http://www.nlm.nih.gov/mesh/trees2008/C08.pdf

[25] MEDLINE Indexing: Online Training Course: Subheadings. National Library of Medicine. Retrieved at March 12, 2012. http://www.nlm.nih.gov/bsd/indexing/training/SUB_010.htm

[26] International Health Terminology Standards Development Organisation. History of SNOMED CT. Retrieved at: January 10, 2012. http://www.ihtsdo.org/snomed-ct/history0/

[27] Zhang, M. Patrick, J., and Truran, D. Inference of a SNOMED CT data model. *Bridging the Digital Divide: Clinician, Consumer and Computer*. (Eds.) J. Westbrook, J. Callen and G. Margelis. Health Informatics Society of Australia (HISA): Victoria, 2006.

[29] Tim Benson. *Principles of Health Interoperability HL7 and SNOMED.* Springer, London, pages 186-202, 2010.

[30] Flora Lum. Structure of SNOMED CT. Created at: April 2005. Retrieved at: April 2, 2012. http://pdffinder.net/Structure-of-SNOMED-CT.html

[31] U.S. National Library of Medicine. UMLS: FAQs: SNOMED CT. National Institute of Health. Retrieved at: February 7, 2012. http://www.nlm.nih.gov/research/umls/Snomed/snomed_faq.html

[32] Klaus A. Kuhn, James R. Warren, and Tze-Yun Leong. *MEDINFO 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics.* IOS Press,U, pages 640-644, September 15, 2007.

[33] Michael Schopen and Stuart J Nelson. ICD-10 and the Unified Medical Language System (UMLS). Meeting of heads of WHO Collaborating Centres for the Classification of Diseases. World Health Organisation, Brisbane, Queensland, Australia, October 14-19th, 2002.

[34] Andy Sager. ICD-10: Dual Coding vs. Double Coding. Created at: August 16, 2011. Retrieved at: April 01, 2012. https://3mhealthinformation.wordpress.com/2011/08/16/icd-10-dual-coding-vs-double-coding/

[35] ICD10 Watch. Dual Coding: QuadraMed coding system supports ICD-9 and ICD-10 code sets. Created at: August 24, 2011. Retrieved at: April 01, 2012. http://www.icd10watch.com/blog/dual-coding-quadramed-coding-system-supports-icd-9-and-icd-10-code-sets

[36] F. Rendy Vogenberg. *Understanding Pharmacy Reimbursement.* American Society of Health-System Pharmacists, pages 56-57, January, 2005.

[37] Josef Ingenerf and Thomas Diedrich. Notwendigkeit und Funktionalität eines Terminologieservers in der Medizin*. Künstliche Intelligenz,* pages 6-14 , March, 1997.

[38] Open Galen. Open Galen. Retrieved at: Februar 10, 2012.
http://www.opengalen.org/background/background0.html

[39] Alan R. Aronson. MetaMap: Mapping Text to the UMLS Metathesaurus. Created at: July 14, 2006. http://skr.nlm.nih.gov/papers/references/metamap06.pdf

[40] Alan R. Aronson. The effect of textual variation on concept based information retrieval. In *Proceedings of the 1996 AMIA Annual Fall Symposium*, pages 373–377. Hanley and Belfus, Philadelphia, PA, 1996.

[41] Alan r. Arson. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. In: *Proc. Of the Annual AMIA Aymposium,* pages 17-21, 2001.

[42] Alan R. Aronson. MetaMap Candidat Retrieval. Created at: March 29, 1991. Modified at: July 13, 2001. Retrieved at: February 10, 2012.
http://skr.nlm.nih.gov/papers/references/metamap06.pdf

[43] Divita, G., Tse, T. and Roth, L.. Failure analysis of MetaMap Transfer (MMTx). In: Fieschi, M., Coiera, E., Li, Y.C.J. (Eds.), MEDINFO 04, IOS Press, Amsterdam. pages 763-767, 2004.

[44] Angelos Hliaoutakis, Kaliope Zervanou, Europides G. M. Petrakis. The AMTEx approach in the medical document indexing and retrieval application. *Data & Knowledge Engineering archive,* Volume 68 Issue 3, March, 2009.

[45] Clement Jonquet, Nigam H. Shah, and Mark A. Musen. The Open Biomedical Annotator. Summit on Translat Bioinforma. Published online March 1, 2009, pages 56-60.

[46] Thierry Delbecque and Pierre Zweigenbaum. MetaCoDe: A Lightweight UMLS Mapping Tool. *AIME Proceedings of the 11th conference on Artificial Intelligence in Medicine*, pages 242-246, 2007.

[47] Thierry Delbecque and Pierre Zweigenbaum. A GATE plugin for tagging french medical texts with UMLS concepts. *In Proc AMIA Symp*, 2011

[48] Thierry Delbecque and Pierre Zweigenbaum. MetaCoDe: A GATE plugin for tagging Medical Corpora in French with controlled terminologies. Created at October 3, 2011. Modified at: October 10, 2011. Retrieved at: February 10, 2012.

[49] Thierry Delbecque . MetaCoDe V0.2: An UMLS Tagger Plugin in GATE: Project Presentation. Created at: February 20, 2011. Retrieved at: February 10, 2012. http://switch.dl.sourceforge.net/project/metacode/Documents/presentation.pdf

[50] Bhatia N, Shah NH, Rubin DL, Chiang AP, and Musen MA. Technical report: Comparing concept recognizers for ontology-based indexing: MGREP v MetaMap, Stanford University. Created at: September 17, 2008. Retrieved at: February 15, 2012. http://bmir.stanford.edu/file_asset/index.php/1349/BMIR-2008-1332.pdf 2006.

[51] Clement Jonquet, Nigam H. Shah, Cherie H. Youn, Mark A. Musen. NCBO Annotator: Semantic Annotation of Biomedical Data. Stanford Center for Biomedical Informatics. Created at: Spetember 8, 2009. Retrieved at: April 1, 2012. http://www.lirmm.fr/~jonquet/publications/documents/Demo-ISWC09-Jonquet.pdf

[52] Clement Jonquet, Mark A. Musen, Nigam H. Shah. Help will be provided for this task: Ontology-Based Annotator Web Service. Stanford Center for Biomedical Informatics Research. Technical Report, June 27, 2008.

[53] W. R. Hersh, D. H. Hickam, and T. J. Leone. Words, concepts, or both: optimal indexing units for automated information retrieval*. Proc Annu Symp Comput Appl Med Care.* Biomedical Information Communication Center, Oregon Health Sciences University, Portland, pages 664-648, 1992.

[54] W. R. Hersh and L. C. Donohoe. SAPHIRE International: a tool for cross-language information retrieval. *Proc AMIA Symp.* Division of Medical Informatics and Outcomes Research, School of Medicine, Oregon Health Sciences University, Portland, USA, pages 673-677, 1998.

[55] Lawrence H. Reeve. Semantic Annotation and Summarization of Biomedical Text. A Dissertation, Drexel University, July 2007.

[56] Kim GR, Aronson AR, Mork JG, Cohen BA, and Lehmann CU. Application of a Medical Text Indexer to an online dermatology atlas. *Stud Health Technol Inform.* Division of Health Sciences Informatics, Johns Hopkins University School of Medicine, Baltimore, pages 287-291, 2004.

[57] U.S. National Library of Medicine. MetaMap Indexing Algorithm. Last Modified at: March 16, 2004. Retrieved at: April 10, 2012. http://ii.nlm.nih.gov/MTI/mmi.shtml

[58] Clifford W. Gay, Mehmet Kayaalp, and Alan R. Aronson. Semi-Automatic Indexing of Full Text Biomedical Articles. *AMIA Annu Symp Proc*, pages 271-275, 2005.

[59] Alan R. Aronson, James G. Mork, Clifford W. Gay, Susanne M. Humphrey, and Willie J. Rogers. The NLM indexing initiative's medical text indexer. Stud. Health Technol. Inform. Pages 268-272, 2004.

[60] U.S. National Library of Medicine. Medical Trxt Indexer (MTI). Last Modified at: November 30, 2011. Retrieved at: April 10, 2012. http://ii.nlm.nih.gov/mti.shtml

[61] Qinghua Zou, Wesley W. Chu, Craig Morioka, Gregory H. Leazer, and Hooshang Kangarloo. IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing. *Proc AMIA Symp*, 2003

[62] Y. L. Zieman and H. L. Bleich. Conceptual mapping of user's queries to medical subject headings. *Proc AMIA Annu Fall Symp*. Pages 519–522, 1997.

[63] Center for Advanced Information Management. MedLEE: A Medical Language Extraction and Encoding System. Technology Fact Sheet. Created at: February 8, 2006. Retrieved at: April 2, 2012. http://www.cat.columbia.edu/pdfs/MedLEE_2006.pdf

[64] Carol Friedman, Pauline Kra, Hong Yu, Michael Krauthammer, and Andrey Rzhetsky. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *In: ISMB Supplement of Bioinformatics*. Pages 74-82, 2001.

[65] Alan R Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances*. J Am Med Inform Assoc* , pages 229-236, 2010.

[66] A. Hliaoutakis, K. Zervanou, G, E. E. Milios. Automatic document indexing in large medical collections. *In Proceedings of the international workshop on Healthcare information and knowledge management*, 2006.

[67] Alan (Lan) R. Aronson, François-Michel Lang, James (Jim) G. Mork, Willie Rogers, an Sonya E. Shooshan. MetaMap Portal. Created at: January 12, 2012. Retrieved at: April 11, 2012. http://metamap.nlm.nih.gov/datafilebuilder.pdf

[68] H. J. Lowe, I. Antipov, W. Hersh, C.A. Smith, an M. Mailhot. *Methods of information in Medicine.* Automated Semantic Indexing of Omage Reports to Support Retrieval of Medical Images in the Multimedia Electronic Medical Record. Schattauer Verlagsgesellschaft mbH, 1999.

[69] F. M. Carrero, J. C. Cortizo Pérez, J. M. Gómez Hidalgo, Testing Concept Indexing in Crosslingual Medical Text Classification. Third IEEE International Conference on Digital Information Management (ICDIM), London, UK. Pages 512-519, November 13-16, 2008.

[70] Shah NH, Bhatia N, Jonquet CM, Rubin DL, Chiang AP, and Musen MA. *Comparison of Concept Recognizers for building the Open Biomedical Annotator.* BMC Bioinformatics, 10(Suppl 9): S14, September 2009.

[71] U.S. National Library of Medicine. Biomedical Research & Informatics: UMLS: Implementation Resources: MetaMap. National Institute of Health. Created at: May 21, 2010. Modified at: September 8, 2011. Retrieved at: April 25, 2012.
http://www.nlm.nih.gov/research/umls/implementation_resources/metamap.html