The use of the medical ontology for a semantic-based fusion system in Biomedical Informatics

Application to Alzheimer Disease

Roxana - Oana TEODORESCU
Computer Science Faculty
”Politehnica” University of Timisoara, Romania
roxana.teodorescu@cs.upt.ro

Cosmin CERNAZANU-GLAVAN
Computer Science Faculty,
”Politehnica” University of Timisoara, Romania
cosmin.cernazanu@ac.upt.ro

Vladimir - Ioan CRETU
Computer Science Faculty,
”Politehnica” University of Timisoara, Romania
vladimir.cretu@cs.upt.ro

Daniel RACOCEANU
IPAL CNRS, Singapore (CNRS, NUS, I2R, UJF)
French National Research Center
A/Prof. University of Besançon, France

Abstract

The Unified Medical Language System (UMLS)[1] *offers the possibility to use annotated medical terms for Computer Aided Diagnoses System (CADS). We present a new semantic fusion system, based on UMLS. This fusion system has applications on a CADS that diagnoses neurodegenerative diseases. Since the UMLS Metathesaurus contains a huge amount of data, classification and extraction of the data we use is necessary. For this purpose, we use a feedforward neural network which is capable of training the negative patterns as well as the positive ones. At the semantic level we generate a three-layered network structure, which gives us the possibility of adding medical knowledge in order to cluster the data and prepare it for the fusion process.*

1. Introduction

In a world where the population is aging continuously, dementia diseases are the leading cause of death. Alzheimer’s disease (AD) alone is the fifth leading cause of death in USA, according to the Alzheimer’s Association [1]. It is currently estimated that each 71 seconds one American develops Alzheimer and by the 2050 this frequency is estimated to reach 33 seconds. As at the moment only symptomatic therapy is available for these diseases, therefore a system that can identify these diseases and also offer early diagnoses is needed. University College of London researchers have reached a 96% accuracy on detecting AD [2]. Early detection is vital [2] because a treatment can prolong the development of the disease by preventing the deterioration of the brain. Early detection constitutes in our case the main application of the proposed fusion system. We choose this approach as more information can be extracted using different types of sources. In our case we use several types of images (fMRI and SPECT) because they give us heterogeneous and homogeneous information. In order to give a unitary result on the processed images, the homogeneous information is fused.

The fusion process takes place at the semantic level which, in our case, is represented by the Semantic Network from UMLS. The Unified Medical Language System (UMLS) was developed by the National Library of Medicine (NLM) and is a multi-language, multi-purpose database containing a huge thesaurus called Metathesaurus [10] [5]. A *concept (CUI)* represents a medical term that in our case is extracted as a feature from the image level and is mapped to the UMLS Metathesaurus, using medical-based rules. Each concept can be linked to one or more high-level *semantic type* in the Semantic Network (SN). The semantic types have links between them, called *relations*. The semantic types and their relations constitute the Semantic Network (SN). This is the highest level of data in UMLS, allowing us to better synthesize the primary information. The SN level corresponds to the ontology information used in the fusion process.

Due to the fact that these sematic types, 135 in total, have 54 kinds of relations among them and that some of them are more relevant for our study than others, we implement a system that classifies these types according to their medical

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

importance for our study. This classification helps identify the semantic types and which of the three layers it belongs to: *anatomical*, *physiological* or *disorder*. The three categories are defined and introduced by Dr. Bodenreider [3]. We use these categories, obtained from our neural network, as a knowledge base (for the training phase). The choice of the neural network is based on the fact that *the generalization* is made automatically as a result of their structure. At this level we fuse the data and prepare it for the decision step where the diagnoses is given. The feedback from the medical doctors offer us the possibility to modify the parameters and the functions for upgrading the system.

Our approach for a fusion at the semantic level is new as compared with other fusion systems like *KnowBaSICS-M* system, although using UMLS as well, its fusion produces an image as a result and this fusion is preformed at a lower level. A semantic fusion system of CT images based on CBIR is presented by Miao and Miao in [6]. This system uses a weighted complex similarity retrieval algorithm and takes into account knowledge support, but it retrieves images and is not based on the semantic axes of the features. A multimodal data fusion between SPECT and MRI, based this time on a geometrical model, is given by Montagner in [7]. We extract the geometrical elements in order to complete the information at the Meta level, but the fusion does not depend on this feature.

In Section 2 we present our system at the ontology level. Section 3 completes the theory with several preliminary results and we present our conclusion together with future perspectives for our system in Section 4 .

2. Using UMLS in our system

We use UMLS in our system because it is complete, contains medical knowledge coming from more vocabulary, and is permanently updated. UMLS combines a large variety of ontologies and terminologies in the Specialist Lexicon, the Metathesaurus and Semantic Network [4]. UMLS is in principle a knowledge base meant to define and classify numerous themes in medicine and to be used in medical-related fields.

For each medical term annotated in UMLS at the Metathesaurus level there is a Concept Unique Identifier (CUI) that corresponds to that term in all the languages contained in the databases part of this system. The same identifier contains all the synonyms possible for that term, as well as a definition for it.

The Semantic Network level is represented by the *semantic type*, the atoms in this network that have a unique identifier attached. The SN is a hierarchic structure where the atoms are linked together by *semantic relations*. These types contain information on the tree hierarchy that they are part of, their description, as well as definitions. This network is meant to categorize all the concepts from the Metathesaurus, by using their relations [9]. In our approach, the goal is to map the extracted features onto the Metathesaurus as concepts. Using these concepts, we access the Semantic Network (SN) level, where the categorization is made using the Semantic Relations, thereby adding the medical knowledge (Fig. 1).

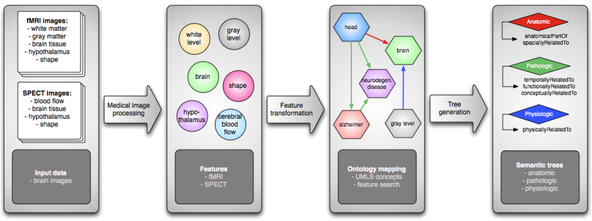

Figure 1. The use of ontology in our system

2.1. The Metathesaurus use in the fusion system

The Metathesaurus is reached by overcoming the semantic gap - the difference between the two imaging techniques (fMRI and SPECT) on one hand and the visual elements and the medical knowledge on the other hand. It is composed of a set of files containing the terms from various thesauri, classified and coded, organized by concept or meaning [8]. This thesaurus contains alternative names of the same concept and also includes the relationships among concepts. These concepts and their CUIs constitute the first step in what we use from the UMLS, as we take these concepts and using the relationships with the Semantic Network we make the transition to another level.

At the Metathesaurus(Meta) level we use the weight functions for taking into account the importance of the concepts in the system and the mapping into the SN.

The parameter for choosing the relation in the Meta - *metaP* is applied in order to differenciate between the relations that exist in the Meta level linked to a concept.

There are six possible relationships in the Meta that we take into account : *SIB* - *SIB*ling; *RO* - *R*elation *O*the associative relation; *RN* - *R*elation *N*arrower than ... ; *RB* - *R*elation *B*roader than ...; *CHD* - *CH*il*D*; *PAR* - *PAR*ent - in reverse order of their importance degree. For expressing that importance degree, we took a function that is capable of adding value to the concepts for the Meta (equation 1), taking into account the relations that derive from the concept being analyzed.

$$y_{metaP} = f(metaR) = metaR/6 \tag{1}$$

where

$$\begin{array}{l} metaR = \{0, RO, SIB, RN, RB, CHD, PAR\} \\ f(0) = 0 \\ f(PAR) = f(CHD) \\ f(PAR) > f(RB) \\ f(RB) = f(RN) \\ f(RN) > f(RO) > f(SIB) \end{array} \tag{2}$$

For the metaP parameter we take into account the relationship types in their order of importance so that we achieve the maximum value of this parameter for the PAR/CHD relations 1.

For mapping the concepts from the Meta into the SN, UMLS uses relations. This translation from the Meta level to SN is specified by the *mapP* parameter. This parameter takes into account the semantic type(s) associated to each concept.

$$y_{mapP} = f(STY) = STY/32 \tag{3}$$

where

$$\begin{array}{l} y_{mapP} > 1 \\ STY \in \{ \begin{array}{l} anatomicalstructure, bodypart, cell, \\ tissue, physiologicalfunction, etc. \end{array} \} \end{array} \tag{4}$$

The condition takes into account the fact that to each concept corresponds to at least one sematic type. We take into account 32 semantic types for our system: 11 for anatomy, 12 for disorders and 9 for physiology.

At the Meta level the *final value* of the mapped features will be given by the val_{CUI_i} from equation 5.

$$val_{CUI_i} = val_{imgExtr} + \sum_{i \in \mathrm{Rel}_{CUI_i}} metaP_i \tag{5}$$

where the value attached to the concept i identified with CUI_i is computed from $val_{imgExtr}$ represents the value of the primary feature extracted from the image level and $metaP_i$ represents the value from equation 1 applied for all the relations in the Meta that correspond to the concept i.

Taking a look at the SN level we can see the link between the two levels in the UMLS, but we must also take into account that important information could be lost at the transition between them. That is the reason why we apply functions that take into account the source of the information and its trajectory until the highest level of abstraction.

2.2. The Sematic level and the information granularity

The Semantic Network is designed to help researchers develop computer aided systems and offer an annotated semantic ontology. The semantic types with their relationships offer a network that can be used for information extraction, as well as for classification. In our case we identify the semantic types related to the concepts identified from the Meta, where we apply the rules of importance as parameters. We use the files generated by MetamorphoSys for extracting the semantic types used and their relationships. The medical knowledge we add to this information is represented as parametrical functions.

At the SN level we construct medical-based semantical trees based on anatomy, disorders and physiology, as shown in the last step from Fig.1. The classifications for the semantic types that are part of these trees are made by Dr. Bodenreider in [3]. His classification is improved by applying clusters in a neural network that learns the semantic types that are used by our system and their importance as well. These semantic types and the relations among them have different importance levels. The importance levels are expressed as parametric functions. For this purpose, we use the semantic network tables.

Using his classification we have given each category an importance degree contained in the *treeP* parameter. This parameter takes into account the type of tree we are dealing with, as well as the image type that generates this semantic type. This is due to the fact that certain image types have a better result on anatomy, whereas other types give better results on pathology. We also take into account the type of the relation that is established between the sematic types - whether it is inside the tree (intra-tree) or between trees (inter-tree).

$$\begin{array}{l} y_{treeP} = f(treeType, tree\mathrm{Rel}, imgType) \\ = imgType - abs(1 - \sqrt{treeType}) \\ +treeType * \sin(tree\mathrm{Rel}) \end{array} \tag{6}$$

where

$$\begin{array}{l} treeType = \{ANA, PHYS, DISO\} \\ tree\mathrm{Rel} = \{0, \mathrm{int}ra, \mathrm{int}er\} \\ imgType = \{fMRI, SPECT\} \\ f(ANA, tree\mathrm{Rel}, fMRI) \\ > f(ANA, tree\mathrm{Rel}, SPECT) \\ f(PHYS, tree\mathrm{Rel}, fMRI) \\ < f(PHYS, tree\mathrm{Rel}, SPECT) \end{array} \tag{7}$$

The *treeType* parameter gives the importance of the trees and has to take one of the values in this set as each concept corresponds to at least one semantic type. On the other hand, the *treeRel* parameter can take the 0 value, since it may occur that there are no relations linked to a specific semantic type. For the *imgType* parameter the values correspond to the image types processed and their values are meant to respect the rules from (7).

At the SN level, we have the final value computed in equation 8 and its value given by val_{STY_i} which includes the value of val_{CUI_i} from 5 for all the mappings in the SN of the CUI_i. The values of the intra and inter-tree relations are given by $SREL_{STY_i}$ and correspond to all the trees that the semantic type i appertains - $TreeType_{STY_i}$. These parameters are computed in equations 3 for the mapping in the SN, 6 for the tree appurtenance and respectively **??** for the relations on the SN.

$$\begin{array}{l} val_{STY_i} = val_{CUI_i} * \sum\limits_{i \in STY_{CUI_i}} mapP_i + \\ \sum\limits_{j \in TreeType_{STY_i}} \sum\limits_{k \in SREL_{STY_i}} treeP_j * (SN_{\mathrm{Rel}})_k \end{array} \tag{8}$$

This final value is the one that will be managed in the fusion process at the end of the ontology phase, giving the medical importance and also making the translation to the semantic level. Note that we use the neural network at the SN stage.

All these parameters give the importance of the data at the end of this stage, before the fusion of the two images, making those that can be identified in more than one layer on the final tree network more importance than those that are just in one layer. Also the data that comes from a specific layer that has more importance from the medical point of view will have a greater value than the others. We must now find the balance between these parameters and their importance in the final value of the data. In a preliminary phase we prepare the data for the fusion system by defining the three layers network. The neural networks are used for classifying the semantic types into the network and giving to each of them an importance degree.

This allows us to use the created setting and to activate only the atoms forming the network that are found in the processed images. These atoms receive their values according to the functions in 8 for the importance degree computation.

3. Preliminary results

We have performed several tests at the SN level using the data provided from fMRI and SPECT extraction so far. As the mapping step into the Meta concepts give us one or more concepts, at the SN we will have one or more semantic type for each concept from Meta. The semantic types identified are then classified into the tree categories. This classification is possible using a feedforward neural network that is capable of giving negative examples as well as positive ones.

From training this network we obtain a feedback on the data range used in the neurodegenerative diseases. Once the parameters have been balanced and normalized we have the possibility of using this data and just map the extracted results on the parameters already computed.

Our classification method gives us results of 56.69% and 100%, for the STY "qualitative concept" that is not classified in one of the trees, and for "congenital abnormality" classified in the DISO tree respectively. These results offer us the degree of association of each STY to the tree.

4. Conclusion and Future work

From our point of view, the semantic tree generation of data represents the most important step in this system. Together with the rule generation, this step introduces the medical knowledge but also transfers the data from the medium level- Meta- to the high level -SN.

Also the fact that we use UMLS for this purpose, a vast data pool from which we extract what we need, offers us the possibility to rely on several generated ontologies. The SN classified by one of its authors also helps us in this step. The importance of this ontology-based system is given by its multiple applications. From the UMLS point of view it represents the natural next step towards a clustering at the semantic level, as it is based on a classification of semantic types.

The *semantic tree concept* included in the three-layered network system, with the appropriate functions for including the medical importance, can be applied not only for other diseases, but also in other ontology-data integration systems as well.

We will test the network for clustering the data according to all the parameters introduced. For now we have classification results on the semantic type and the parameter y_{mapP} (equ. 3). After testing the clustering with all the parameters we will perform classification on the final value given by the function V_{STY} (equ. 8). We are confident that this will provide us with an environment able to take into account all the variables presented in this article, based on medical knowledge.

References

[1] A. Association. 2008 alzheimer's disease facts and figures. *Elsevier - Alzheimer's and Dementia*, 4(2):110–133, February 2008.

[2] N. BBC. Computers' spot alzheimer's fast. News article MMVII, BBC, February 2008.

[3] O. Bodenreier, B. Smith, A. Kumar, and A. Burgun. Investigating a subsumption in snomed-ct: An exploration into large description logic-based biomedical technologies. *Artificial Intelligence in Medicine*, 39(3):183–195, 2007.

[4] M. Kohler. *UMLS for Information Extraction*. PhD thesis, Vienna University of Technology, May 2007. Institute of Software Technology & Interactive Systems -Information Engineering Group.

[5] C. Lacoste, J.-P. Chevallet, J.-H. Lim, D. L. T. Hoang, X. Wei, D. Racoceanu, R. Teodorescu, and N. Vuillenemot. Inter-media concept-based medical image indexing and retrieval with umls at ipal. *Lecture Notes in Computer Science, Evaluation of Multilingual and Multi-modal Information Retrieval*, 4730/2007:694–701, 2007.

[6] Y. Miao and Y. Miao. The research of semantic content applied to image fusion. *Proceedings of the 32-nd Applied Image Pattern Recognition Workshop*, (8072300):125–130, Oct. 2003.

[7] J. Montagner, V. Barra, and J.-Y. Boire. Synthesis of a functional information with anatomical landmarks by multiresolution fusion of brain images. *Engineering in Medicine and Biology Society 27th Anual Conference*, (6):6547–6550, September 2005. ISBN: 0-7803-8741-4; 17281770 [PubMed - in process].

[8] N. L. of Medicine. Unified medical language system - section 2 - metathesaurus. Documentation 2008AA, United states National Library of Medicine, May 2008.

[9] N. L. of Medicine. Unified medical language system - section 3 - semnatic network. Documentation 2008AA, US National Library of Medicine, May 2008.

[10] D. Racoceanu, C. Lacoste, R. Teodorescu, and N. Vuillemenot. A semantic fusion approach between medical images and reports using umls. *Lecture Notes in Computer Science, (Eds.): Asian Information Retrieval Symposium*, 4182/2006:460475, octobre 20 2006.