Medic-Us: Advanced Social Networking for Intelligent Medical Services and Diagnosis

Gandhi Hernández Chan[1], Alejandro Molina Villegas[1], Mario Chirinos Colunga[2], Oscar S. Siordia[2], and Alejandro Rodríguez González[3]

[1] Conacyt - Centro de Investigación en Ciencias de la Información Geoespacial
[2] Centro de Investigación en Ciencias de la Información Geoespacial
[3] Centro de Tecnología Biomédica, Universidad Politécnica de Madrid

Abstract. Health services are on the top priorities for society, but up to now we have fail in make it universal all around the world. Nowadays information technologies, specially social networks have demonstrated its usefulness in different areas. This article describes the design and development of a social network platform focused on the physician-patient and physician-physician interactions, in order to achieve better and faster diagnosis. Like other social networks or social media tools, it focus on the collaboration among its members. This collaboration is improved with the help of paradigms as Collaborative Intelligence and Wisdom of the Crowd. We called this platform Medic-Us highlighting the collaborative practice among the practitioners, and the interaction with patients. This document describes the different modules of Medic-Us, the social network environment, medical consult service, information retrieval, and a trainer module for the medicine students.

Keywords: Social Networks · Decision Support System · Semantic Web · Diagnostic System · Collaborative Intelligence.

1 System in a nutshell

This section has the aim of present the Medic-Us system components and the relation among them. The system is a web 2.0 platform so, it was developed as a social network taking into account that the main features of that kind of software are concepts such as Collective Intelligence and Wisdom Of the Crowd. This is because through the information exchange among the physicians the system pretends to offer a better service. In section §2 we mention how these concepts and technologies brings benefits to Medic-Us. In section §4 we present the architecture of the system, which is organized in four layers (presentation, services, business logic and data). The presentation layer corresponds to the Social Network Environment described in section §5 which includes the use of NLP techniques for the implementation of a sentiment analysis module based on the users comments. The NLP was also used for the knowledge enhancement described in section §7 analyzing the content of medical knowledge sources such as MedlinePlus. For the data layer we decided to use Semantic Web technologies (ontologies) as knowledge representation that was useful to establish the relation among diseases and other elements that physician use for diagnosis process such as signs and test. This is described in section §6. As it can be observed, Medic-Us system has many related components and each one is needed in order to obtain the expected results

2 Introduction

Web 2.0 applications have promoted and improved interaction and collaboration between people all around the world. Some authors mention that online interaction can help people by motivating them, as well as by decreasing their isolation feelings [17]. From this new ways of interactions arise new concepts to describe the way in which this new interactions result on valuable data or information for solving different kind of problems. The first of these concepts, called Collaborative Intelligence (CI) [15], refers to the ability of a group to solve more problems and give better solutions by working together than by summing up their individual contributions. Another term is called Wisdom Of the Crowds (WOC), where in adequate circumstances groups can become extremely intelligent, and even smarter than the smartest person within the group, and as more people get involved the results improve [38]. The WOC is defined as a process that takes into account the collective opinion of a group of individuals rather than a single expert to answer any question. However, the results are not always the best ones because of the biases as result of selection of the crowd.

In [1] it is mentioned that WOC is one of the most important concepts when building collaborative platforms. This is because there are different Web 2.0 applications, such as wikis, forums, blogs and podcast, as well as other websites that allow their users to write comments. This kind of applications is largely described by [3], [31]. Is interesting to notice that a great amount of these collaborative websites are related to health content. This situation has allowed authors such as Eysenbach [8] to address the concept of e-health, in order to refer to "an emerging field in the intersection of medical informatics, public health and business, referring to health services and information delivered or enhanced through the Internet and related technologies. In a broader sense, this term characterizes not only a technical development, but also a state-of-mind, a way of thinking, an attitude, and a commitment for networked, global thinking, to improve health care locally, regionally, and worldwide by using information and communication technology". Examples of e-health sources are Google health cards and Wikipedia, popular sites that medical students and professionals use as informational resources for medical subjects. However, some of the inconvenient that people may face when using Google for medical diagnosis purposes is that less than the 40% of the results contain high quality information [18]. Something similar happens when using Wikipedia, which has neither the most accurate, complete and reliable information from the pharmacological point of view. That is why it is highly recommended to students and professionals to consult more reliable sources [21].

The ability of a physician to diagnose a patient's condition depends on several factors such as education, training, experience and available resources. One of this resources is the collaborative network of peers that the physician has. To take advantage of this network (Collective Intelligence) and the individual experience of a large set of physicians (WOC), we have developed Medic-Us, a collaborative web environment to provide health services. It's main feature is that the health information is given only by medical experts consulting a Clinical Decision Support System (CDSS). In Medic-Us we also take care of the communication channels, this is because we know that only a physician can treat a patient, that the patient's information must be kept in secret, and that it is risky that none expert advise a patient about health issues. So communication among the patients is not allowed.

One of the motivations to create Medic-Us is a previous study [14] where it is presented that collective intelligence can achieve better diagnosis than individual physicians. Having a social network like Medic-Us provide some extra benefits apart form the CI and WOC, e.g. can provide specialized diagnosis in communities where there is not always a specialized physician, or for people with mobility problems how needs medical diagnosis without going to a hospital. Further, it can reduce the waiting time in the hospitals for a diagnosis.

This document is organized as follows: §3 presents the state of the art focusing on collaborative Web sites with medical content, section §4 describes the architecture of the system, section §5 presents the social network environment, its main features and functionality, in section §6 we mention how the knowledge representation was built, §7 describes the knowledge acquisition process, section §8 presents the virtual medical office, and how it works with the CDSS, section §9 presents the medical trainer module, and finally, section §10 presents our conclusions and future work.

3 Related Work

In [19] Kamel and Wheeler argue that Web 2.0 applications and technologies can become great enablers for health and healthcare professionals due to the fact that through them people can use collective intelligence in democratic ways for generating new knowledge, interacting and sharing resources, experiences and responsibilities. As they mention, nowadays there are websites such as the British Medical Journal and others [4] were people, including health professionals and students, interact in different ways such as creating private groups or sending private and public messages where they share different kind of information. To improve the experience some sites offer the possibility of filtering the people you interact with, by dividing users into categories, such as verified practitioners (people with proved medical license or a GMC certificate) or normal users. Is wroth notice that these authors also argue that the possibilities that these technologies offer can be highly improved by combining them with some

[4] patients.co.uk, SurWiky, HealthyPlace, PatientOption and doc2doc

of the characteristics of the Semantic Web or Web 3.0. As an example, they mention that human-computer interfaces can be simplified by enabling users to be in more control over how information is accessed, as well as to provide better search and information retrieval algorithms.

In [9] Giustini presents examples of Social Network Sites (SNS) where digital interfaces are used for store, organize, share and discus medical diagnosis. One of the examples he mentions is the Ves Dimov's Clinical Cases and Images website, were several clinical cases, from different specialties are described and evaluated, sometimes supported by images. In order to do this, the web page includes a list of the probable diagnosis that were generated according to a clinical tests results, as well as an explanation of the reasons for the given diagnosis. Another example mentioned by the author is a collaborative medical *wiki* named Ganfyd, which is commonly used as a diagnosis reference by medical professionals and invited non-medical experts. This site enables its users to share their knowledge in a *wiki* format. The content is supervised by qualified physicians with a certificate from UK General Medical Council, or a valid account at the doctors.net.uk or ausdoctors.net.

Sandars and Schroter [32] conducted a semi structured online questionnaire survey to students and qualified medical practitioners on the British Medical Association database. The main finding showed that there is a high familiarity with Web 2.0 technologies, but few practical use. The surveyed population related this situation with the lack of training on how to use these technologies in relation to educational purposes. In addition, they also mentioned that some other barriers that need to be faced when using online technologies were the concerns about the quality of the resources, the lack of time, and technological issues that difficult access.

In [10] Giustini states that the Semantic Web, or Web 3.0, could become a powerful tool for medicine by allowing medical practitioners, bioinformatics and researchers to locate, process and extract larger amounts of data from disparate systems. As he mentions, Web 3.0 will also allow doctors to develop a more personalized healthcare system, and will promote the reduction of medical treatments costs by making the search for health information more efficient and responsive to patients' needs.

In [4] Boulos et al. explains that their easy use, and the capability to interact, collaborate, and share information in a free or low cost way is what has contributed to social networking proliferation between clinical practice and education environments. Nonetheless, these characteristics also represent some of their biggest problems due to the fact that almost everyone can alter, edit or contribute these collaborative documents without control. However, these authors also assert that even when there are different problems that can result on serious quality issues, collaboration behind this kind of websites most of the times follows a Darwinian type of process, in which their content tends to be improved. In addition, they also mention that there are specific ways to assure the veracity and quality of the information, through the establishment of a monitoring and moderating system in a closed environment.

Gruber [11] claims that social networking sites and the Semantic Web could be combined in order to generate Collective Knowledge Systems applications. The author defines this kind of applications as human-computer systems in which machines enable the collection and harvesting of large amount of human-generated knowledge. The three main parts of a Collective Knowledge System are: a social system supported by technology, a search engine, and users. The advantages that these collaborative intelligence based systems have, in comparison to traditional systems, is that the contents are user-generated. They also have human-machine synergy, which means that the provided information will be more accurate to the users need because the range of coverage provided by these systems is wider and it is based on official and other reliable sources.

In relation to Ontology construction, Zhdanova [43] explains the added value that community driven portals can have by presenting a framework that promotes the construction of collaborative Ontology portals in which users are the ones who define the content structure and the ways in which the content is managed. The author argues that the current portals are usually comprehensive, but also limited due to the fact that they do not allow a complete participation of the users, because that they cannot modify the portals structures in order to solve their needs. She also claims that a larger degree of portal flexibility and adaptation to member's real demands can be achieved by updating the existing community web portals through semantic web technologies. Despite this argument, she mentions that giving users the ability to add new attributes to the Ontology also represents a risks in terms of obtaining undesirable results such as a bad structure, unreliability, inefficiency and/or redundant activities. To solve this problems she states that it is possible to create and structure-generic Ontologies supported by the community and/or to support the development of domain-dependent Ontologies created by collaboratively by end users and domain experts. However, in order to build community-drive Ontologies there must be a consensus process, in which an individual creates an ontology item(s) and/or data that is relevant for him/her. After that, the community members discover the relevance of the created item, and finally everything returns to the first step. Additional examples of ontology management methodologies, such as IST EU projects, DIP, SEKT, KnowledgeWeb, SWWS, Esperonto, and WonderWeb, are also reported at her work.

4 Medic-Us

Medic-Us was built as a CI and WOC system based on a social network and supported by a clinical decision support system. Figure 1 shows the architecture of the system with its components. As it can be seen, the architecture is divided in four layers presentation, services and business logic.

4.1 Presentation layer

This layer corresponds to the social web environment, where the patients can access to the virtual medical office and select their from a list. Next they choose a physician they want to consult with; i.e. the practitioner that will review their symptoms (Consult Process). As a requirement, the doctor and the patient have to be connected in the social network. After the consult process, the doctor will receive the list of symptoms, the patient data, and a list of probable diagnosis. With this information, and using the social web, the doctor contacts the patient to give him a diagnosis, a prognosis and a possible course of treatment.

4.2 Services layer

This layer contains a two web services that were built using the RestFul architectural style and XML as information format. One service manage the symptoms and other manage the diagnosis. The symptoms symptoms web services transport the information provided by the user to the inference engine in the next layer. The diagnosis web services send the results form the inference engine and the patients information to the patient's preferred physician. This service layer allows the user to intact throw a web portal or a mobile app.

4.3 Business logic layer

This layer contains the inference engine that infer probable diagnosis from the provided symptoms, the information stored in the knowledge data base and a file with the heuristics to follow. The diagnostic list and the patient's data will be sent to the database in order to keep track of the patient's health record, and also sent to the physician via e-mail. The physician will also receive an mail alert.

4.4 Data layer

This layer contains three elements, Knowledge database, Heuristics file and Patients database. The knowledge database has the information about the diseases and its related symptoms. The heuristics file, establishes the relation between the diseases and its symptoms in order to obtain a possible diagnosis. The patients database, has the information about the patients and their health records. This is useful for the physicians because they can use this information in order to know the current patient's health and it's evolution.

5 Social Network Environment

The social network takes into account three main roles, physicians, patients and administrator. As a medical expert or as a patient the first step to enter the social network is to register. Any one can register as a patient without restriction. The registration process is similar to many others, the user have to write his/her name, surname, email, and other general information. In the case of physicians the registration includes a validation process of their professional license number (PLN) in order to ensure that they are actual medical experts. For this the registration consist of two parts. In the first part practitioners write their general information including their professional license number. In the second part, the website administrator has to validate the PLN. To do this, he consult official channels, if the information provided is valid then the administrator grants the permission to belong to the community.

As it is common in other social networks, members can create groups. In Medic-Us this is restricted only to physicians. They can look for other physicians and send them a connection request, once the request is accepted, they can create groups based on their specialty or any other interest. The figure 2 shows the website panel from the doctor's point of view. The site is in Spanish because the project was designed and built for being used in Latin America countries, especially in México.

5.1 Sentiment Analysis of Physician's Reviews

In order to enhance the user experience and the collaboration between users (patients and physicians) an amiability ranking was included. The amiability ranking is based on the polarity of the comments made by users in the platform and it is used to create suggestions for the interaction among users. Thanks

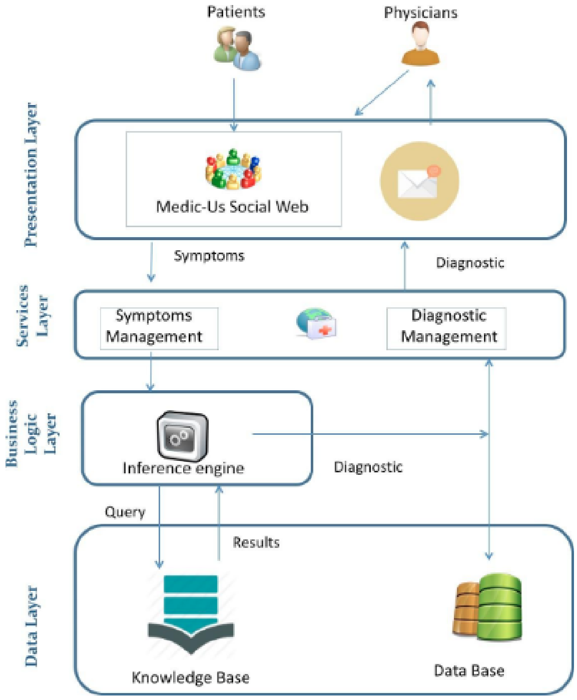


Fig. 1. Medic-Us Architecture.

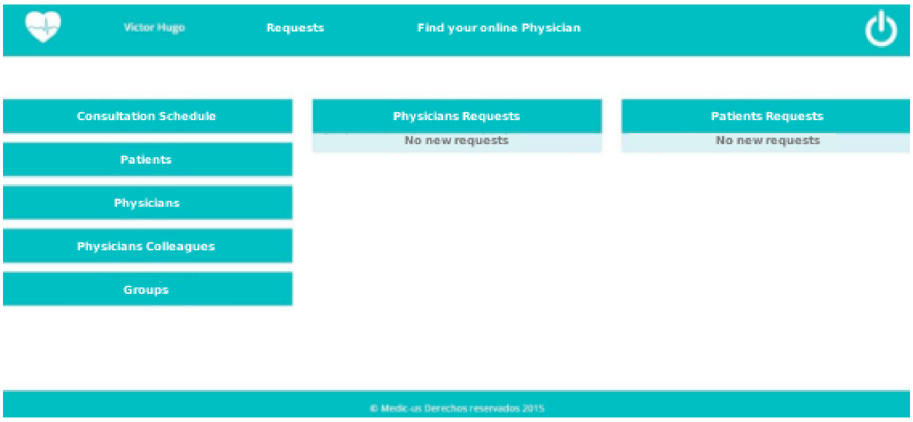
to this module, on the one hand, the physicians are able to know the patients mood before starting an interaction, which results useful as previous information for the diagnosis. On the other hand, the patients will be able to know the reputation of the available physicians based on the polarity detected on the messages exchanged with other patients. The sentiment Analysis process is based in NLP techniques presented in [41] where the authors determine the polarity of a given text using a Support Vector Machine.

6 Knowledge Representation

We built the knowledge representation for this research based on the work presented in [29], which is a set of ontologies that were designed to be directly used with Diagnostic Decision Support Systems. This set of ontologies use SNOMED-CT [35] as its supported terminology. Based on this situation, we associated each of the concepts presented at the ontologies to a SNOMED-CT code in order to make them efficient for reasoning and inference possible diseases through the construction of differential diagnosis.

The decision of using this ontology was taken based on two main reasons. First of all, due to the fact that by reusing it we would later be able to make comparisons based on one same criterion. On the second place, it needs to be mentioned that, even when most works address the theme of methodologies as ontology integration, such as in [6] and [27]; in our case applying these same techniques was not possible because, even when we had the required data, it was necessary to build a root ontology in order to define the relationships between concepts. The medical knowledge representation can be seen in Figure 5.

For the development of the set of the ontologies protégé software was used. Figure 3 shows a view of the set of ontologies The left side shows the hierarchy of the set of ontologies that support the CDSS including DO (Disease Ontology)



No new requests

No new requests

© Medic-us Derechos reservados 2015

Fig. 2. Medic-Us Web Site Panel.

with the information about the diseases such as its name and Snomed-CT code, DRO (Drugs Ontology). This ontology has the information of the drugs related with each disease through the **can_occur_with** relation. This is because some drugs come with side effects that can be presented as signs or even as a disease, DTO (Diagnostic Test Ontology) has the information about the diagnostic test that are related with a particular disease in order to get a better diagnosis, and SO (Signs Ontology), that has the information about the signs of each disease. The relations among the ontologies are used to create the model of each disease, as we can observe in Figure 4. The right side shows the relations between DO and the rest of the ontologies. This is because a Disease Model is consists of signs, diagnostic tests, and even other diseases.

Figure 4 shows an example of a disease model. The final part of each line has a code and the name of the ontology that the item belongs to. So the first item is I25374005 and belongs to the DO. The I letter means that it refers to an Item in the ontology, and the code corresponds to Gastroenteritis disorder in Snomed-CT terminology. Then, the items two to five and the last one refers to signs that defines the Gastroenteritis, so I68962001 is Muscle Pain, I25064002 refers to Headache, I43724002 refers to Chill, I267060006 refers to Diarrhea symptom, I386661006 refers to Fever, and I16932000 refers to Nausea and vomiting.

7 Knowledge Enhancement

Because the ontology in [29] had only the 30 most common diseases on family medicine, it was necessary to extend the ontology with more knowledge in order to be used with more clinical cases and to be able to attend more patients. We choose MedlinePlus [24] as source of information to extend our ontology over

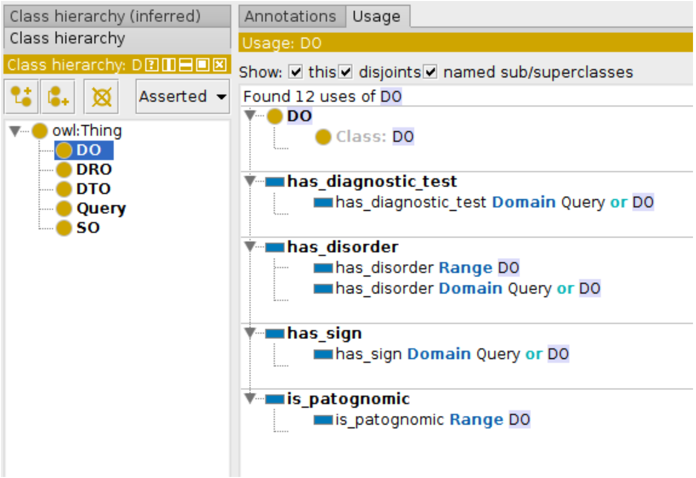


Fig. 3. Ontologies explorer view.

other efforts such as [42], [39], [12] and [2] because in contrast with MedlinePlus they are focused on more specific medical areas, and do not address the extraction of the basic clinical terms used in a diagnostic process and from collaborative sources and textual databases such as Wikipedia and Freebase which also contain valuable knowledge, but the reliability and completeness of their information is questionable [30].

MedlinePlus is an online free information service provided by the US National Library of Medicine, which is considered the world's largest medical library. It provides reliable and up-to-date health information from the National Institutes of Health and other trusted sources on over 1000 diseases and conditions, extensive information on prescription and nonprescription drugs, and links to thousands of clinical trials. The data provided for each disease may vary, but it usually includes a description of the disease, causes, symptoms, exams and tests, and treatment.

The information from MedlinePlus was extracted using the process described in [30], a prototype capable of crawling webpages in order to extract all relevant diagnosis-related content (symptoms, sings and diagnostic tests), an then apply a named-entity recognition approach to extract all relevant terms based on MetaMap. The output of the process is a list of diagnosis-related terms for each disease.

7.1 MedlinePlus Extraction

The process to extract the information in MedlinePlus described in [30] consist of three steps: a) Medical Text Extraction and NLP Procedures (MTENP).

```
<rdf:Description rdf:about="file:///C:/gandhi/kbmedglobal/disont.owl#I25374005">
    <ddxont:has_sign rdf:resource="file:///C:/gandhi/kbmedglobal/signs.owl#I68962001"/>
    <ddxont:has_sign rdf:resource="file:///C:/gandhi/kbmedglobal/signs.owl#I25064002"/>
    <ddxont:has_sign rdf:resource="file:///C:/gandhi/kbmedglobal/signs.owl#I43724002"/>
    <ddxont:has_sign rdf:resource="file:///C:/gandhi/kbmedglobal/signs.owl#I267060006"/>
    <ddxont:has_disorder rdf:resource="file:///C:/gandhi/kbmedglobal/disont.owl#I16932000"/>
    <ddxont:has_sign rdf:resource="file:///C:/gandhi/kbmedglobal/signs.owl#I386661006"/>
</rdf:Description>
```

Fig. 4. Example of data structure for a disease model.

b) Validation Terms Extraction Procedure (VTE Procedure) and c) Validation Procedure (TV Procedure).

Medical Text Extraction and NLP Procedures (MTENP) This step comprises the MetaMap filter and produces a list of annotated medical terms. For doing this, the URL of a selected disease is sent to the MTE module which applies a web scraping procedure and extracts the text of the relevant sections of the page, and then, applies MetaMap to the extracted text. The filter process results in a list of relevant annotated medical terms based on the semantic types from SNOMED-CT.

When used as a compositional terminology SNOMED-CT can accurately represent 92.3% of the terms used commonly in medical problem lists. Improvements to synonymy and adding missing modifiers would lead to greater coverage of common problem statements. Health care organizations should be encouraged and provided incentives to begin adopting SNOMED CT to drive their decision-support applications [7].

Validation Terms Extraction Procedure (VTE Procedure) This module improves the terms produced by MetaMap by obtaining medical terms from other sources of different type. Official sources includes ICD9CM, ICD10CM and Mesh, research sources include CCSO Sings and Symptoms Ontology, TM Sings and Symptoms Ontology (TM SSO) and Symptoms Ontology as well as collaborative sources including Wikipedia and Freebase, MedicineNet.

Validation Procedure (TV Procedure) This module is responsible for analyzing the terms provided by the MTENP procedure to ensure they match VTE-provided terms. If the TV procedure finds a match, the term is returned as a valid diagnostic term. The validation process attempts to find a match between a given term t obtained from the list provided by MTENP and a matching

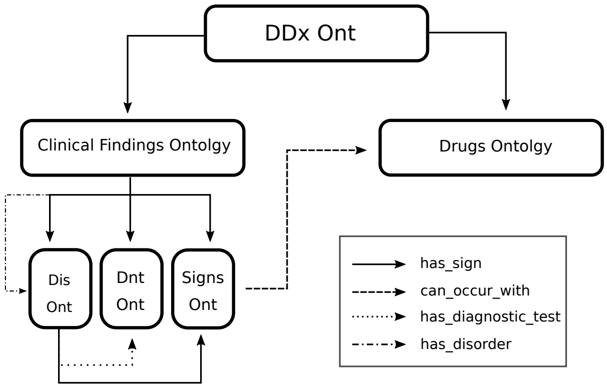


Fig. 5. Knowledge Representation.

term *mt* obtained from the list provided by VTE. If a matching is found, it is assumed that the term t is a valid diagnostic term and it is added to the final list of results. We used the list of results as a clinical-terms list with the name and code of the diseases and its related symptoms and tests, and then we used the Jena API [22] to add the list to the ontology. This improved list of symptoms and their context are a valuable resource for summarization and named entities recognition.

7.2 Automatic Summarization of Medical Literature

In order to efficiently process the massive volumes of information filtered from Medline Plus we applied a Summarization technique called Sentence Compression. The main idea is to link a complete text with its shorter representation that contains the important parts and some of the context. The compressed texts versions of Medline Plus texts serve both as a succinct extract of evidence about a particular term and as the input of a machine learning method for Symptoms Extraction described in §7.3.

Sentence compression establishes a bridge between extraction and abstraction since it realizes a fine-grained processing, creating a primary form of paraphrase: a telegraphic version. The term *sentence compression* was used for the first time in [20]. The authors define Sentence Compression as follows: let φ be a sentence as a sequence of n words, $\varphi = (w_1, \ldots, w_n)$. An algorithm must eliminate words so that the remaining sequence is a compression of the original text (to change the order of the words is not allowed).

In [23], authors propose a dynamic programming algorithm that decides, for each word, if the sentence gets a better score by keeping it o eliminating it. The score is a linear function based on the features of the sentence and its compressions, which weights are calculated using a training corpus.

Our approach for Sentence compression is based on more recent studies that add two important aspects to the sentence compression task: 1) it is necessary to consider the context of the sentence instead of processing it in an isolate way and 2) it is more natural to eliminate fragments from the sentence than isolate words. So the elimination of discourse structures instead of propositions has been explored and authors argued that, even if automatic discourse analysis at the document level is still a challenge, discourse segmentation at the sentence level is a realistic alternative to the sentence compression [26, 36]. Considering the two mentioned important aspects, we have used the approach of Sentence compression by discourse units elimination, originally proposed in [25, 26].

The approach for the module of Automatic Summarization of Medline Plus texts was to generate compression candidates (CC) by deletion of some discourse segments from the original sentence. Let S be a sequence of k discourse segments: $S = (s_1, s_2, \ldots, s_k)$. A compression candidate, CC_i, is a subsequence of S that preserves the original order of the segments. The original sentence always forms a candidate, i.e. $CC_0 = S$, this is convenient because sometimes there is no shorter grammatical version of the sentence, especially in short sentences that conform one single EDU. Since we do not consider the empty subsequence as a candidate, there are $2^k - 1$ candidates. We used Statistical Language Modeling as a technique to assign a probability to a sequence of words. The probabilities in a Language Model (LM) were estimated counting sequences from Medline so we based our estimations using large corpora and interpolation methods. Following the method described in [25], we used a big corpus to obtain the sequence counts and a LM interpolation based on Jelinek-Mercer smoothing [5]. In a LM, the maximum likelihood estimate of a sequence is interpolated with the smoothed lower-order distribution. We used the Language Modeling Toolkit SRILM [37] to score the segment likelihood probability assuming that good compression candidates must have a high probability as sequences in a LM.

Finally, the compressed versions of Medline Plus texts containing symptoms were used as the input of a machine learning classifier capable of recognize similar context in unseen texts to detect new symptoms in order to enrich the ontology.

7.3 Named Entities Recognition for Symptoms Extraction

Text Mining have raised as promising solutions for one of the most challenging aims of the digital age: transforming data into insights.

The problem addressed in this part is the automatic detection and classification of entity names in domain specific documents. This process is known as Named Entity Recognition (NER) and systems capable of high performance on this task are desirable because NER precedes other relevant NLP tasks including Information Extraction. So, the performance of a entity recognizer affects directly the performance of complex Text Mining systems. Therefore, NER is considered the cornerstone for some ambitious projects and that is why has been an active research area for some years and has been recently applied in many fields going from Medicine [40] and Chemistry [28] to Geology [34] and History [33].

It has been observed that successful algorithms for Information Extraction in news suffer a significant drop in performance when they are applied to Medical-Biology documents. Along *BioCreAtIvE* campaigns, results for advanced tasks are significantly lower than reported results using journalist texts, demonstrating the current limitations of text-mining approaches where knowledge extrapolation and interpretation are required [16]. There are some particular aspects of biological discourse to be consider in order to deal with corpora in this field. The terminology is in constantly renewing, full of neologisms. Every day new species appear in scientific papers and many genes and proteins are mentioned for the first time or renew their name in literature. In addition, the interdisciplinary nature of Medicine could complicate assertions about entities. From the point of view of genetics, a microorganism could have a totally different description, and even a different name, than that from biology for instance.

The presented experiments and results focused on discovery of specific medical entities: symptoms. However, it is worth noting that although reported experiments in this work were designed to detect symptoms, the approach presented here is general enough to be applied in any field.

Table 1. Comparison of tools for symptoms names detection in a text about trees.

	Monomial		Binomial		n-gram	
	Precision	Recall	Precision	Recall	Precision	Recall
RegExp (ontology terms)	1.0000	0.6821	1.0000	0.5748	0.9117	0.4033
ANN (perceptron)	0.9120	0.6171	0.9545	0.7608	0.8379	0.5669

Our approach to detect entities is based on artificial neural networks. The main idea is to pass raw text to the input layer of a perceptron. The whole context of the symptom mention is considered in the perceptron training giving as a result a model. The model (trained perceptron) do estimates the probability of a word, more precisely a n-gram, to be a symptom given its textual context.

Table 1 presents the results for the evaluation of two different methods for symptoms detection: a regular expression (Regex) and a Artificial Neural Network model (ANN) trained for medical literature. For the ANN model, we first created a training dataset, the best ANN model obtained with medical literature is the one we use to extract symptoms in order to extend the ontology. Evaluation is composed of three sub-tasks: to seek out one-word symptoms (Monomial), to seek out two-word symptoms (Binomial) and to seek out any length symptoms names (up to seven words).

As expected, extracting names based on RegExp is limited to the original dictionary which is reflected on the low recall. The most significant drawback of Regex is that it does not recognize subtle differences in spelling. It simply does not match terms if they are not written exactly as they appear in the regular expression.

To get more flexibility in the detection of symptoms, ANN presents a more stable approach. It is capable of finding terms that do not have the same orthography of the lexicon. In that sense, it does not depend on a dictionary once the model is trained which is very convenient for neologisms discovering.

According to the results we concluded that the best strategy for symptoms names detection is to combine both methods RegExp and ANN. The final strategy of the Knowledge Enhancement module is to use the initial ontology for the RegExp method and the ANN to discover new terms in the corpus.

8 Virtual Medical Office

The Medic-Us project has some common services of social networks, but the most important service is a Virtual Medical Office (VMO). This service makes the difference with other social networks that address medical issues. The VMO is based on a CDSS that uses the set of ontologies presented in §6. For using the VMO a patient needs to be user of Medic-Us and connected to a physician. When a patient uses the VMO, First he login into Medic-Us, gets into the VMO module and select his symptoms from a list, e.g. fever, headache, chest pain, vomit, etc. The VMO page can be seen in figure 6. After the patient has listed all his symptoms he selects the physician he wants to consult and the physician receives the list of symptoms, the probable diagnosis and the patient data as an e-mail, figure 7.

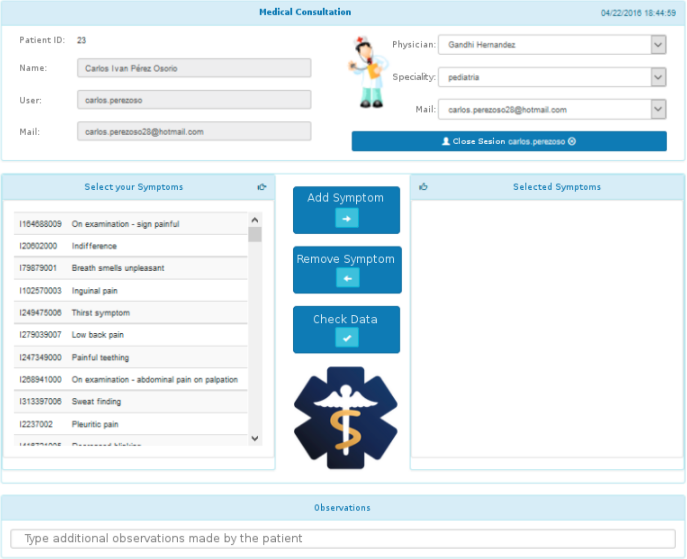
Finally, the doctor will contact the patient through the social web to explain the diagnosis, the prognosis and treatment. In case the patient doesn't need to visit the doctor personally, the doctor has the option to create and send a digital prescription to the patient.

9 Medical Trainer

The Medical Trainer is the second most important module in Medic-Us Social Web. It is also based on the set of ontologies presented in the Knowledge Representation Section. For this module we used the cosine similarity metric to find closeness between diseases. Then, a disease and its findings (signs/symptoms) are shown to a medical student who must select the correct answer from a list of four possible diseases. Once the student has finished, the system grades the test and then their teacher will give them a feedback. A desktop version of this trainer was presented in [13] and the architecture of the Medical Trainer is shown in figure 8.

10 Conclusions

The data deluge of on-line medical information represents a great opportunity for computer scientist to mine all of the knowledge in intelligent ways. However,



Medical Consultation

04/22/2016 18:44:59

Patient ID: 23

Name: Carlos Ivan Pérez Osorio

User: carlos.perezoso

Mail: carlos.perezoso28@hotmail.com

Physician: Gandhi Hernandez

Speciality: pediatria

Mail: carlos.perezoso28@hotmail.com

Select your Symptoms

I164688009	On examination - sign painful
I20602000	Indifference
I79879001	Breath smells unpleasant
I102570003	Inguinal pain
I249475006	Thirst symptom
I279039007	Low back pain
I247349000	Painful teething
I268941000	On examination - abdominal pain on palpation
I313397006	Sweat finding
I2237002	Pleuritic pain

Selected Symptoms

Observations

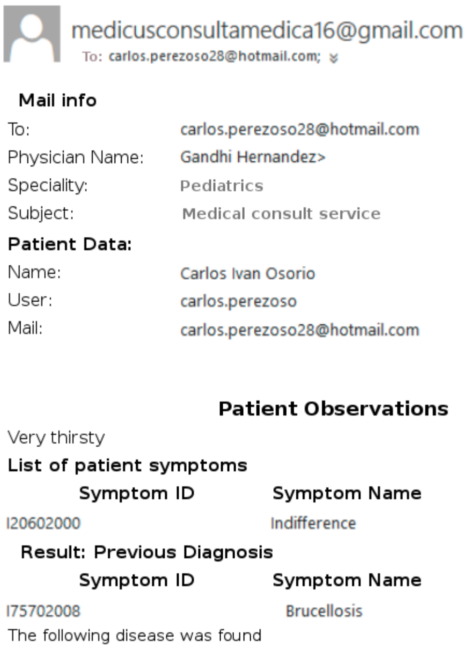
Type additional observations made by the patient

Fig. 6. Virtual Medical Office Panel.

while some medical data sources follow structured templates or international standards, some other remain unstructured or partially structured. We can access now to thousands of repositories of ontologies, data tables, images, videos and texts. For these reasons, sophisticated approaches from Data Mining and Natural Language Processing have proved their benefits for unstructured data processing of literature in the domain of Medicine.

In Medicine, to obtain a correct diagnosis is paramount. Moreover, deal with patients information is a sensitive matter that requires expert supervision. We are convinced that a follow-up could be even more robust when more expert point of views are included. With this in mind, we designed Medic-Us, a collaborative web environment to provide health services. It's main feature is that the health information is given only by medical experts consulting a Clinical Decision Support System (CDSS). In Medic-Us, the initial diagnosis is automatically and then a refined diagnosis can be made using the Wisdom Of the Crowds paradigm, only that this time we include exclusively Physicians. Medic-Us, unlike other platforms, implements a CDSS that uses a modular ontology as knowledge representation. Regarding the security and privacy aspects we can mention that Medic-Us include all the necessary validations of practitioners and cryptography based management of patients information access.

With this platform we are addressing some common gaps that many web pages and social networks with medical content has. The first one is the com-



medicusconsultamedica16@gmail.com
To: carlos.perezoso28@hotmail.com;

Mail info

To:	carlos.perezoso28@hotmail.com
Physician Name:	Gandhi Hernandez>
Speciality:	Pediatrics
Subject:	Medical consult service

Patient Data:

Name:	Carlos Ivan Osorio
User:	carlos.perezoso
Mail:	carlos.perezoso28@hotmail.com

Patient Observations

Very thirsty

List of patient symptoms

Symptom ID	Symptom Name
I20602000	Indifference

Result: Previous Diagnosis

Symptom ID	Symptom Name
I75702008	Brucellosis

The following disease was found

Fig. 7. Mail with the consultation data.

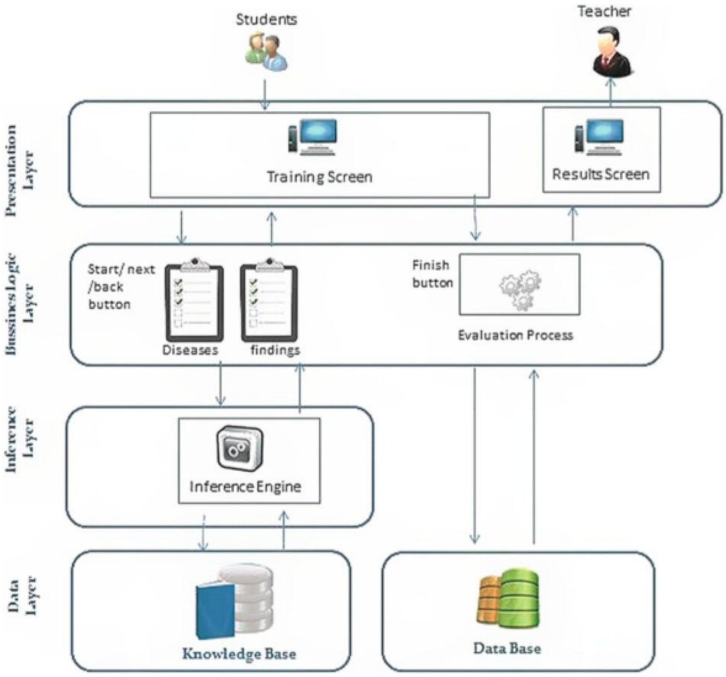


Fig. 8. Architecture of Medical Trainer.

munication between medical experts and patients. The main difference among Medic-Us and other platforms is that Medic-Us allows the patient to communicate with some specific physicians through making social groups, but avoid the communication among patients. The second one is the communication among medical experts in order to interchange information in different formats such as text, images and video. This is in order to enhance the medical service and the diagnostic precision and accuracy through the Collective Intelligence reached with the participation of a group of physicians trying to solve an specific case. The third one is that Medic-Us functionality is based on the use of Semantic Web technologies such as a Clinical Decision Support System that use a medical domain ontology as knowledge base.

As part of the future work Medic-Us will include a module that allow a better recommendation of the physicians. The idea is that depending on the initial diagnosis of the patient, the system will recommend specialist based on probabilistic methods. We plan also to include more sophisticated methods for the initial diagnosis based on semantic relations metrics between ontology nodes. Also, as future work we pretend to use NLP techniques in order to obtain new information that could feed the knowledge base based on experts comments and opinions in the social network and compared with clinical terms of Snomed-CT and MedlinePlus.

References

1. Alag, S.: Collective intelligence in action. Manning Publications Co. (2008)
2. do Amaral, M.B., Roberts, A., Rector, A.L.: Nlp techniques associated with the opengalen ontology for semi-automatic textual extraction of medical knowledge: abstracting and mapping equivalent linguistic and logical constructs. In: Proceedings of the AMIA Symposium. p. 76. American Medical Informatics Association (2000)
3. Barsky, E.: Introducing web 2.0: weblogs and podcasting for health librarians. Journal of the Canadian Health Libraries Association/Journal de l'Association des bibliothèques de la santé du Canada **27**(2), 33–34 (2006)
4. Boulos, M.N.K., Maramba, I., Wheeler, S.: Wikis, blogs and podcasts: a new generation of web-based tools for virtual collaborative clinical practice and education. BMC medical education **6**(1), 41 (2006)
5. Chen, S., Goodman, J.: An empirical study of smoothing techniques for language modeling. Computer Speech & Language **13**(4), 359–393 (1999)
6. Corcho, O., Fernández-López, M., Gómez-Pérez, A.: Methodologies, tools and languages for building ontologies. where is their meeting point? Data & knowledge engineering **46**(1), 41–64 (2003)
7. Elkin, P.L., Brown, S.H., Husser, C.S., Bauer, B.A., Wahner-Roedler, D., Rosenbloom, S.T., Speroff, T.: Evaluation of the content coverage of snomed ct: ability of snomed clinical terms to represent clinical problem lists. In: Mayo Clinic Proceedings. vol. 81, pp. 741–748. Elsevier (2006)
8. Eysenbach, G.: What is e-health? J Med Internet Res (Jun 2001)
9. Giustini, D.: How web 2.0 is changing medicine. British Medical Journal Publishing Group (2006)
10. Giustini, D.: Web 3.0 and medicine. British Medical Journal Publishing Group (2007)
11. Gruber, T.: Collective knowledge systems: Where the social web meets the semantic web. Web semantics: science, services and agents on the World Wide Web **6**(1), 4–13 (2008)
12. Hahn, U., Romacker, M., Schulz, S.: Medsyndikate—a natural language system for the extraction of medical information from findings reports. International journal of medical informatics **67**(1-3), 63–74 (2002)
13. Hernandez-Chan, G.S., Ceh-Varela, E.E., Cervera-Evia, G., Quijano-Aban, V.: Using semantic technologies for an intelligent medical trainer. In: International Symposium on Intelligent Computing Systems. pp. 74–82. Springer (2016)
14. Hernández-Chan, G.S., Ceh-Varela, E.E., Sanchez-Cervantes, J.L., Villanueva-Escalante, M., Rodríguez-González, A., Pérez-Gallardo, Y.: Collective intelligence in medical diagnosis systems: a case study. Computers in biology and medicine **74**, 45–53 (2016)
15. Heylighen, F.: Collective intelligence and its implementation on the web: algorithms to develop a collective mental map. Computational & Mathematical Organization Theory **5**(3), 253–280 (1999)
16. Hirschman, L., Yeh, A., Blaschke, C., Valencia, A.: Overview of biocreative: critical assessment of information extraction for biology. BMC bioinformatics **6**(Suppl 1), S1 (2005)
17. Israel, B.A.: Social networks and social support: implications for natural helper and community level interventions. Health education quarterly **12**(1), 65–80 (1985)

18. Judd, T., Kennedy, G.: Expediency-based practice? medical students' reliance on google and wikipedia for biomedical inquiries. British Journal of Educational Technology **42**(2), 351–360 (2011)
19. Kamel Boulos, M.N., Wheeler, S.: The emerging web 2.0 social software: an enabling suite of sociable technologies in health and health care education. Health Information & Libraries Journal **24**(1), 2–23 (2007)
20. Knight, K., Marcu, D.: Summarization beyond sentence extraction: a probabilistic approach to sentence compression. Artificial Intelligence **139**(1), 91–107 (2002)
21. Lavsa, S.M., Corman, S.L., Culley, C.M., Pummer, T.L.: Reliability of wikipedia as a medication information source for pharmacy students. Currents in Pharmacy Teaching and Learning **3**(2), 154–158 (2011)
22. McBride, B.: Jena: Implementing the rdf model and syntax specification. In: Proceedings of the Second International Conference on Semantic Web-Volume 40. pp. 23–28. CEUR-WS. org (2001)
23. McDonald, R.: Discriminative sentence compression with soft syntactic evidence. In: Proceedings of EACL. vol. 6, pp. 297–304 (2006)
24. Miller, N., Lacroix, E.M., Backus, J.E.: Medlineplus: building and maintaining the national library of medicine's consumer health web service. Bulletin of the Medical Library Association **88**(1), 11 (2000)
25. Molina, A.: Compresión automática de frases: un estudio hacia la generación de resúmenes en espanol. Inteligencia Artificial **16**(51), 41–62 (2013)
26. Molina, A., Torres-Moreno, J.M., SanJuan, E., Da Cunha, I., Martínez, G.E.S.: Discursive sentence compression. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 394–407. Springer (2013)
27. Pinto, H.S., Gómez-Pérez, A., Martins, J.P.: Some issues on ontology integration. IJCAI and the Scandinavian AI Societies. CEUR Workshop Proceedings (1999)
28. Rocktäschel, T., Weidlich, M., Leser, U.: Chemspot: a hybrid system for chemical named entity recognition. Bioinformatics **28**(12), 1633–1640 (2012)
29. Rodriguez-Gonzalez, A., Hernandez-Chan, G., Colomo-Palacios, R., Miguel Gomez-Berbis, J., Garcia-Crespo, A., Alor-Hernandez, G., Valencia-Garcia, R.: Towards an ontology to support semantics enabled diagnostic decision support systems. Current Bioinformatics **7**(3), 234–245 (2012)
30. Rodríguez-González, A., Martínez-Romero, M., Costumero, R., Wilkinson, M.D., Menasalvas-Ruiz, E.: Diagnostic knowledge extraction from medlineplus: an application for infectious diseases. In: 9th International Conference on Practical Applications of Computational Biology and Bioinformatics. pp. 79–87. Springer (2015)
31. Rohani, V.A., Hock, O.S.: On social network web sites: definition, features, architectures and analysis tools. Journal of Computer Engineering **1**, 3–11 (2009)
32. Sandars, J., Schroter, S.: Web 2.0 technologies for undergraduate and postgraduate medical education: an online survey. Postgraduate medical journal **83**(986), 759–762 (2007)
33. Smith, D.A., Crane, G.: Disambiguating geographic names in a historical digital library. In: Research and Advanced Technology for Digital Libraries, pp. 127–136. Springer (2001)
34. Sobhana, N., Mitra, P., Ghosh, S.: Conditional random field based named entity recognition in geological text. International Journal of Computer Applications **1**(3), 143–147 (2010)
35. Spackman, K.: Snomed ct style guide: Observables and investigation procedures (laboratory). International Health Terminology Standards Development Organization (2010)

36. Sporleder, C., Lapata, M.: Discourse chunking and its application to sentence compression. In: Conference on Human Language Technology and Empirical Methods in Natural Language Processing. pp. 257–264. ACL (2005)
37. Stolcke, A.: Srilm – an extensible language modeling toolkit. In: Intl. Conf. on Spoken Language Processing. vol. 2, pp. 901–904. Denver (2002)
38. Surowiecki, J.: The wisdom of crowds. Anchor (2005)
39. Tan, K.C., Yu, Q., Heng, C., Lee, T.H.: Evolutionary computing for knowledge discovery in medical diagnosis. Artificial Intelligence in Medicine **27**(2), 129–154 (2003)
40. Tanabe, L., Xie, N., Thom, L.H., Matten, W., Wilbur, W.J.: Genetag: a tagged corpus for gene/protein named entity recognition. BMC bioinformatics **6**(1), 1 (2005)
41. Tellez, E.S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Siordia, O.S., Villaseñor, E.A.: A case study of spanish text transformations for twitter sentiment analysis. Expert Systems with Applications **81**, 457 – 471 (2017). https://doi.org/https://doi.org/10.1016/j.eswa.2017.03.071, http://www.sciencedirect.com/science/article/pii/S0957417417302312
42. Tsumoto, S.: Automated extraction of medical expert system rules from clinical databases based on rough set theory. Information sciences **112**(1-4), 67–84 (1998)
43. Zhdanova, A.V.: Community-driven ontology construction in social networking portals. Web Intelligence and Agent Systems: An International Journal **6**(1), 93–121 (2008)