DOMAIN ONTOLOGY DEVELOPMENT FOR COMMUNICABLE DISEASES

Iti Mathur [1], Hemant Darbari[2] and Nisheeth Joshi[3]

[1,3]Department of Computer Science, Banasthali University, India
[1]mathur_iti@rediffmail.com
[3]nisheeth.joshi@rediffmail.com
[2]Center for Development of Advanced Computing, Pune, Maharashtra, India
[2]darbari@cdac.in

ABSTRACT

Web has become the very first resource to search for any kind of information. With the emergence of semantic web, our search queries have started generating more informed results. Ontologies are at the core of any semantic web application. They help in rapid development of distributed systems by providing information on the fly. This key feature of distribution and sharing of information has made ontologies as a new knowledge representation mechanism. A mechanism which is strongly backed by a sound inference system. In this paper, we shall discuss the development, verification and validation of an ontology in a health domain.

KEYWORDS

Knowledge Representation, Knowledge Engineering, Ontology Development, Health, Communicable Disease.

1. INTRODUCTION

With the emergence of Artificial Intelligence (AI), Knowledge Representation (KR) has been a branch of massive research and development with various KR techniques being developed. In today's web based scenario where information is constantly being shared among different applications, KR techniques such as Scripts, Frames, Semantic Networks does not blend very well. We need a representation technique which can store concepts and relations and which would easily be accessed by different web applications. Ontologies have provided this mechanism. Yahoo1, Amazon2 and Hakia3 are the proof of this. With Yahoo using ontologies for categorization of websites and Amazon using ontologies for categorization of products which are placed for selling by different sellers and Hakia, a search engine using ontologies to implement semantic search for precise retrieval of information.

In recent years, ontologies have gained enormous popularity because it has been backed by World Wide Web Consortium (W3C) as it has developed a framework[1] to encode knowledge on the web pages which shall make the search process easier for web agents which can easily access

[1] http://www.yahoo.com
[2] http://www.amazon.com
[3] http://www.hakia.com

Jan Zizka (Eds) : CCSIT, SIPP, AISC, PDCTA - 2013
pp. 351–360, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3640

Meta information available for the website. This framework has been implemented by ontologies. In fact ontologies are saved in this format and are termed as RDF (Resource Development Framework).

The reason for ontologies being the first choice for semantic web is because they can very easily specify the concepts and unambiguously maintain concept hierarchies. For example, if a user gives a search query as “Manmohan Singh” or “Indian Head Dr. Singh” or “Indian Government Leader” or “Prime Minister of India” to a search engine, then a normal search engine, in some cases would not be able to find the correct results whereas a semantic web based search engine (which is backed by ontologies) can very easily provide the same results to all the search queries. Moreover, if a search query like “plane bomb” is provided to a search engine then it will not be able to relate the two different concepts whereas a semantically informed search engine would relate it with plane bombing and associated treats and would display the results accordingly. This is a clear advantage of semantic search engines (or ontologies) over normal search engines.

This leads us to a question, “What is Ontology”. According to Gruber[2], ontology can be defined as “an explicit specification of conceptualization”. We need ontology because it can easily relate similar concepts and can find associations between similar concepts’ relations and properties, thus reduces the burden of explicitly defining everything.

The rest of the paper is arranged as follows: In section 2 we review the previous work-done in ontology creation and also the work done in health information processing. In section 3 we describe the complete development process of our ontology. In section 4 we evaluate the concepts and properties using a reasoner which implements description logic. Using this reasoner we verify integrity of the concepts, relations and properties. Section 5 concludes the work done.

2. LITERATURE SURVEY

Many researchers have investigated search behaviors of various users for gathering information regarding health. Bhavnani et. al.[3] showed that co-occurrence counts of medical information (like symptoms and disorders) on a web page significantly influence the search behavior. Spink et. al.[4] showed that when health related information is to be searched then users move from general purpose search engines to specialized search engines. Hersh et. al.[5] reviewed medical informatics and information science literature regarding how physicians use IR systems to gain confidence in clinical question answering and decision making. They found that these retrieval systems were inadequate as they retrieved very few relevant articles on a given topic. They did a follow up review on this with another study[6] which showed how students of medicine and nursing use MEDLINE to get information for clinical question-answering. They found out that, with the help of literature searching users were just slightly successful at answering clinical questions. Eastin and Guinsler[7] investigated the relationship between online health information seeking for a symptom or disease and visiting a general practitioner for inquiring the same. They found out that a user’s level of health anxiety moderates the relationship between online health information seeking and health care utilization decisions.

Development of Ontologies has been investigated from different point of views. Hearst[8] started working in this area with concept annotations. Even today, his seminal work on lexico-syntactic patterns is very relevant for annotation based knowledge applications. This work has been refined and reused by several researchers who have applied different approaches. For example, Poesio et. al.[9] extended Hearst patterns for anaphora resolution and used machine learning approaches in identifying patterns. Vashisth et al [10] developed an ontology on human anatomy. This was a heavy weight ontology in which the ontology was divided into four sub-parts. They were:

skeleton structure, nervous system, digestive system, cardiovascular system. In all they created around 600 concepts which have several sub-concepts, relations and properties. Mathur et al [11] developed an ontology on health care services in India. They developed this ontology by collecting health related news articles. Eezioni[12] and Market[13] showed the use of ontologies in Internet by using search engine APIs. Some of the researchers have also applied Lexico-syntactic patterns in the identification of other lexical relations like Carniak and Berland[14] using them in representing part-of relations and Girju and Moldovan[15] representing them for causal relations. Cederberg and Widdows[16] showed how one can improve pattern filtering of Hearst patterns by using Latent Semantic Analysis. Morin and Jacquemin[17] and Ravichandran and Hovy[18] worked on addressing automatic generation of patterns via similarity based approaches in which vectors were formed using same patterns. This approach proved to be more generalized then what was proposed by Hearst.

In literature, we can also find systems which have been developed using these techniques like Buitelaar et. al.[19] showed the use of OntoLT, an ontology learning plugin for Protégé Ontology editor. This system annotated parts of speech chunks, and grammatical relations using a parser. Velardi el. al.[20] showed the use of OntoLearn system where terms were extracted for a domain from a domain-specific textual corpus. This tool became one of the most important tools in automatic creation of ontologies through text. Niang et. al.[21] has shown the requirement of customized development for domain specific ontologies. He argued the ontology development process can be different in different domains and cannot be engineered using the same techniques. Samsfard et. al.[22] showed the development of Persian Word Net using semi automatic processes.

3. PROPOSED WORK

A lot of information is available for health related topics. Most of the websites providing same information for a particular disease or medical procedure use different terminologies. This appears to be a major bottleneck in development of standards and guidelines for annotation of health data. This has inspired us in the development of ontology in health domain which improves the reuse of timely information. Moreover this can map similar concepts and relations available from different sources. For example, a website uses a term HIV+ whereas another website uses AIDS for same information. Although they both are same, search engines would consider them as different search terms which in-turn leads to information related to these terms to be treated as different. Using ontologies we can represent these terms as equivalent terms, thus all the information of one concept can be used by another concept.

3.1 Methodology

Efforts have been put in to develop ontologies as they can store data semantically, which helps in the development of applications for semantic web and the areas where semantic knowledge holds the key value. A key reason of developing ontologies is their ability to learn from real world and to identify the instances and create relations among them. Using ontologies we can generalize health domain terminology and its concepts. As most diseases have same symptoms, it becomes a problem for a normal application to relate which symptom is of which disease, as it creates an ambiguity. Ontological taxonomies are quite useful in disambiguating them. Ontology is a useful tool for developing standards and guidelines for interoperability between various health care information services and heterogeneous web resources. Ontology development is necessarily an iterative process. Each Concept in the Ontology should emulate a real life entity and its relationships in domain of interest. In the following sections we show the development of our ontology.

3.2 Specification

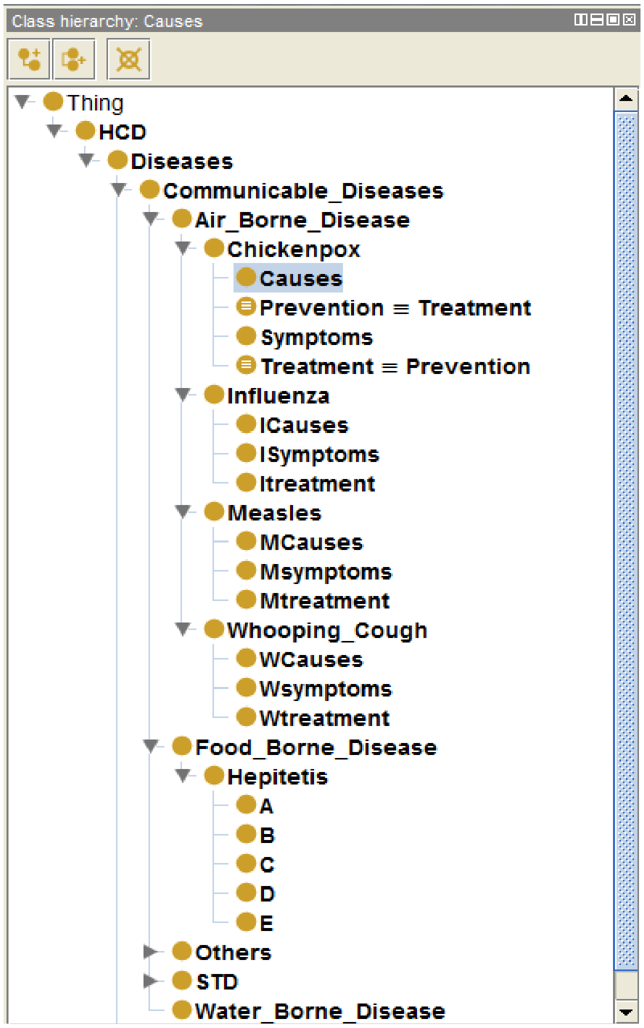
In this phase we collected information regarding various concepts from books, journals and websites. We did a comprehensive study of all the finer points. When we had certain doubts, we contacted experts like doctors, paramedics and medical representatives. They provided us with an insight which helped us clear our confusions. When all the information related to diseases, their possible symptoms and possible cure was collected; then we ascertained the domain and scope of our ontology.

3.3 Conceptualization

We started building our ontology by first implementing the concepts (classes) that were part of our ontology. We created the Human Communicable Disease (HCD) as the super class and all the sub-systems (diseases, symptoms, causes) as its sub classes. While creating different classes, we came across some similar properties and relations which were present in these classes. So, we marked these classes as equivalence classes. Moreover, we also found that some classes did not have common properties. So, these classes were marked as disjoint classes. Figure 1, illustrates these concepts.

After the implementation of classes, we moved onto defining the internal structures of the classes through defining relations, object properties and the relation between two relations. Binary relationship of individuals was implemented through object property. Figure 2 shows the implementation of this in our ontology. The relations used in our ontology were as follows:

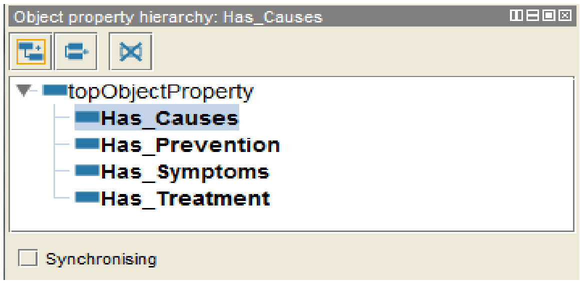
- Antonyms (communicable, Non-communicable)
- Synonyms (flu, Influenza)
- Synonyms (TB, Tuberculosis)
- Is-a (Disease, Health Domain)
- Is-a (Chickenpox, communicable)
- Is-a (Flu, communicable)
- Is-a (Measles, communicable)



Class hierarchy: Causes

- Thing
  - HCD
    - Diseases
      - Communicable_Diseases
        - Air_Borne_Disease
          - Chickenpox
            - Causes
            - Prevention ≡ Treatment
            - Symptoms
            - Treatment ≡ Prevention
          - Influenza
            - ICauses
            - ISymptoms
            - Itreatment
          - Measles
            - MCauses
            - Msymptoms
            - Mtreatment
          - Whooping_Cough
            - WCauses
            - Wsymptoms
            - Wtreatment
        - Food_Borne_Disease
          - Hepitetis
            - A
            - B
            - C
            - D
            - E
        - Others
        - STD
        - Water_Borne_Disease

Figure 1. Structure of Health Ontology



Object property hierarchy: Has_Causes

- topObjectProperty
  - Has_Causes
  - Has_Prevention
  - Has_Symptoms
  - Has_Treatment

□ Synchronising

Figure 2. Object Property of relations

3.5 Creation of Instances

We created the individual instances of the concepts which differentiated between various concepts according to their components. Individual properties and their relations were used for this purpose. Figure 3 shows this instance diagrams.

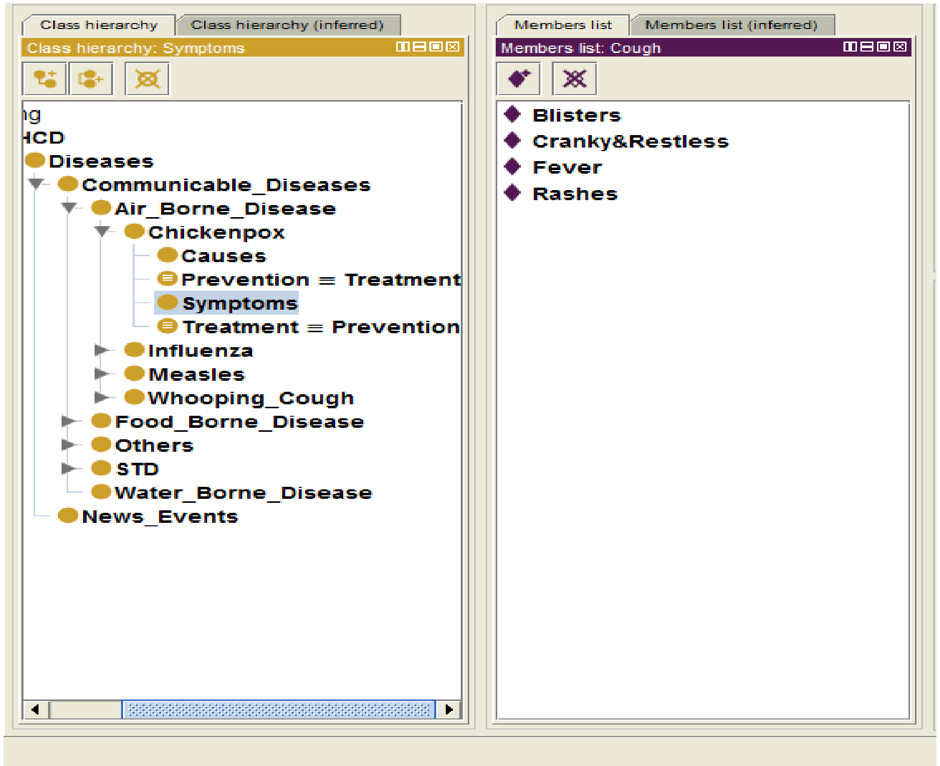


Figure 3. Creation of Instances

3.6 Ontology Visualization

In order to view the relationship between various concepts, relations and their properties we created a graphical view of our ontology. We verified the equivalence relationships between concepts. Figure 4 shows the detailed description of our ontology. Here, the '+' symbol shows that there are concepts and relations which can be expanded. A circle marked rectangles shows the concepts and the diamond marked rectangles show the instance properties. We have marked prevention and treatment as equivalent concepts (classes) in our ontology. This can be seen in the figure as it shows the concepts has equivalent symbol (≡) marked in the circles. Moreover, as seen in the figure, Sneeze is a common symptom for measles and chickenpox. Thus it has an edge connecting both the measles and chickenpox.

4. Evaluation

Since we have developed a resource which will be used by other applications, it was very much necessary to evaluate its correct retrieval. We were also keen on verifying as to whether the ontology has linked the relations and properties correctly or not and whether the relation and properties of equivalent classes are retrieved or not. So, to verify our ontology we used a reasoner based on DL Query which checked each relation, property and in turn the class. We configured the FaCT++[23] reasoner in Protégé for using description logic based queries. When we provided the class name to the reasoner, it retrieved the query in terms of classes, individuals, super class, domain and range. Through the results of the reasoner, we were convinced that ontology is created without any defects as the system was able to retrieve desired classes, super classes and its associated relations and properties. Figure 5 shows the case when treatment was given as an input and it fetched equivalent class, sub-class, super-class etc. In another case we gave causes and executed the query. This fetched all the ancestor super-classes and all the instances of the concept. This is shown in figure 6. Similarly each concept, relation and instance was verified for correct results.

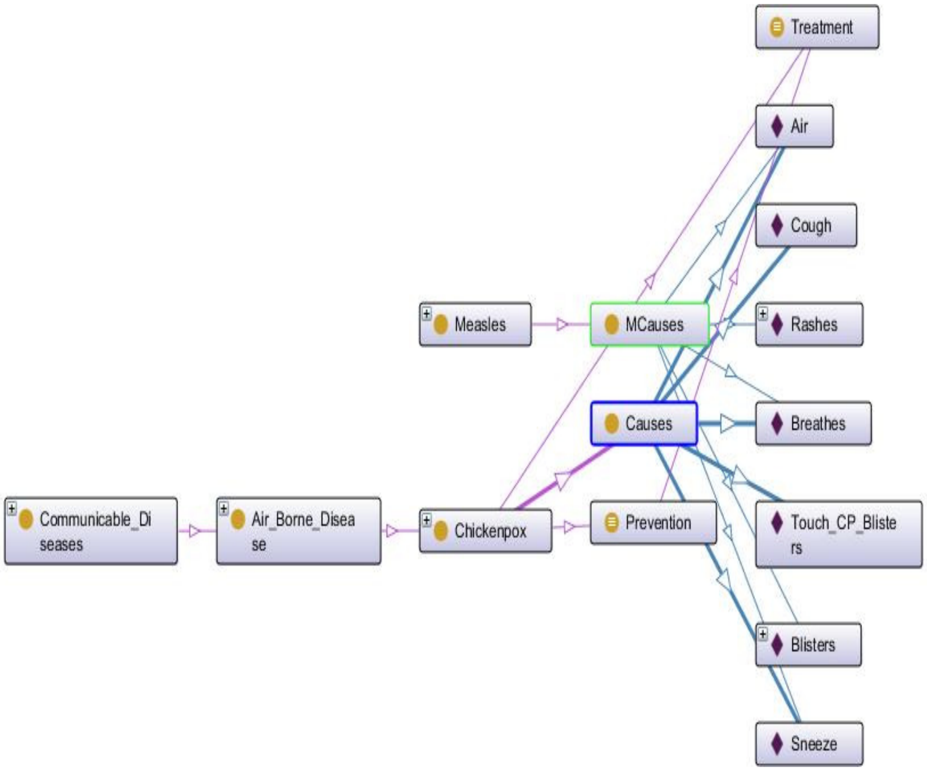
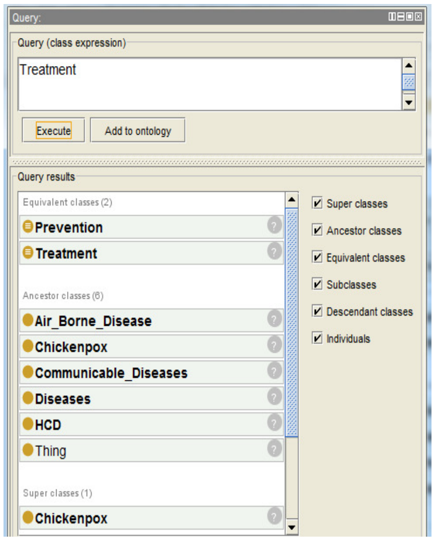


Figure 4. Visualization of Health Ontology



Query:

Query (class expression)

Treatment

Execute | Add to ontology

Query results

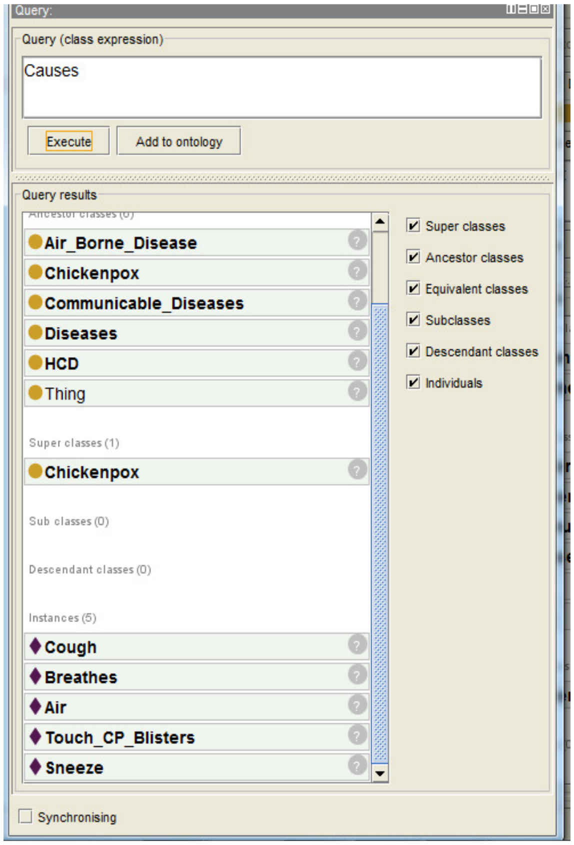
Equivalent classes (2)
- Prevention
- Treatment

Ancestor classes (6)
- Air_Borne_Disease
- Chickenpox
- Communicable_Diseases
- Diseases
- HCD
- Thing

Super classes (1)
- Chickenpox

☑ Super classes
☑ Ancestor classes
☑ Equivalent classes
☑ Subclasses
☑ Descendant classes
☑ Individuals

Figure 5. DL Query for Treatment



Query (class expression): Causes

Ancestor classes (0)
- Air_Borne_Disease
- Chickenpox
- Communicable_Diseases
- Diseases
- HCD
- Thing

Super classes (1)
- Chickenpox

Sub classes (0)

Descendant classes (0)

Instances (5)
- Cough
- Breathes
- Air
- Touch_CP_Blisters
- Sneeze

☑ Super classes
☑ Ancestor classes
☑ Equivalent classes
☑ Subclasses
☑ Descendant classes
☑ Individuals

□ Synchronising

Figure 6. DL Query for Causes

5. CONCLUSION

In this paper we have demonstrated the development process of ontology in health domain. We have discussed various phases in the development of ontology. We have also evaluated our ontology using a reasoner which verified the ontology by executing each class and their properties. As an extension of this work, we would like to enhance the ontology by adding more diseases and its cures. We also wish to connect different components of the ontology with respective web contents and we would develop a front end for this ontology which shall retrieve the results of any disease, its symptoms and would also provide the related information available on the web.

REFERENCES

[1] Brickley, D., Guha, R.V., (1999) "Resource Description Framework (RDF) Schema Specification". http://www.w3.org/TR/PR-rdf-schema.

[2] Gruber, T.R., (1993) "A translation approach to portable ontology specification". Knowledge Acquisition. 5, pp 199–220.

[3] Bhavnani, S.K., Jacob, R.T., Nardine, J., Peck, F.A., (2003) "Exploring the distribution of online healthcare information". ACM SIG Computer Human Interaction Annual Conference. pp 816–817.

[4] Spink, A., Yang, Y., Jansen, J. , Nykanen, P., Lorence, D.P., Ozmutlu, S., Ozmutlu, H.C., (2004) "A study of medical and health queries to web search engines". Health Information and Libraries Journal, 21, pp 44–51.

[5] Hersh, W.R., Hicka, D.H., (1998) "How well do physicians use electronic information retrieval systems? A frame-work for investigation and systematic review". Journal of American Medical Association.

[6] Hersh, W.R., Crabtree, M.R., Hickam, D.H., Sacherek, L., Friedman, C.P., Tidmarsh, P., Mosbaek, C., Kraemer, D., (2002) "Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions". Journal of American Medical Information Association, 9, pp 283–293.

[7] Eastin, M.S., Guinsler, N.M., (2006) "Worried and wired: effects of health anxiety on information-seeking and health care utilization behaviors". Cybernetics and Behavior, 9(4), pp 494–498.

[8] Hearst, M., (1992) "Automatic acquisition of hyponyms from large text corpora". In: 14th International Conference on Computational Linguistics.

[9] Poesio, M., Almuhareb, M., (2005) "Identifying concept attributes using a classifier". In: ACL Workshop on Deep Lexical Acquisition.

[10] Vashisth, A., Mathur, I., & Joshi, N., (2012) "OntoAna: Domain Ontology for Human Anatomy". arXiv preprint arXiv:1208.3802.

[11] Mathur, I., Mathur, S., & Joshi, N., (2011) "Ontology development for health care in India". In Proceedings of the International Conference & Workshop on Emerging Trends in Technology (pp. 715-718). ACM.

[12] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D., Yates, A., (2005) "Web scale information extraction in Know It All". In: 13th World Wide Web Conference.

[13] Markert, K., Modjeska, N., Nissim, M., (2003) "Using the web for normal anaphora resolution". In: EACL Workshop on the Computational Treatment of Anaphora.

[14] Carniak, E., Berland, M., (1999) "Finding parts in a very large corpora". In: 37th Annual Meeting of the Association of Computational Linguistics.

[15] Girju, R., Moldovan, M., (2002) "Text mining for causal relations". In: FLAIRS Conference.

[16] Cederberg S., Widdows, D., (2003) "Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction". In: Conference on Natural Language Learning.

[17] Morin, E., Jacquemin, C., (1999) "Projecting corpus based semantic links on a thesaurus". In: 37th Annual Meeting of the Association of Computational Linguistics.

[18] Ravichandran, D., Hovy. E., (2002) "Learning surface patterns for a question answering system". In: 40th Annual Meeting of the Association of Computational Linguistics.

[19] Buitelaar, P., Olejnik, D., Sintek, M., (2004) "A Protégé plug-in for ontology extraction from text based linguistic analysis". In: 1st European Semantic Web Symposium .

[20] Velardi, P., Navigli, R., Cuchuarelli A., Neri, F., (2005) "Evaluation of OntoLearn, a methodology for automatic population of domain ontologies"

[21] Niang, C., Béatrice, B., Moussa, L., (2010) "Towards tailored domain ontologies". In: 5th International Workshop on Ontology Matching .

[22] Shamsfard, M., Hesabi, A., Fadaei, H., Mansoory, N., Famian, A., Bagherbeigi, S., Fekri, E., Monshizadeh, M., Assi, M., (2010) "Semi Automatic Development of FarsiNet: The Persian WorldNet". In: 5th International Conference on Global WorldNet (2010).

[23] Tsarkov, D., Horrocks, I., (2006) "Fact++ Description Logic Reasoner: System Discription". In: International Joint Conference on Automated Reasoning. Lecture Notes on Artificial Intelligence. 4130, pp 292-297. Springer.

AUTHORS

Mrs. Iti Mathur is an Assistant Professor at Banasthali University. Her primary area of research is Computational Semantics and Ontological Engineering. Besides this she is also involved in the development of MT engines for English to Indian Languages. She is one of the experts empanelled with TDIL Programme, Department of Electronics and Information Technology (DeitY), Govt. of India, a premier organization which foresees Language Technology Funding and Research in India. She has several publications in various journals and conferences and also serves on the Programme Committees and Editorial Boards of several conferences and journals.

Dr. Hemant Darbari, Executive Director of the Centre for Development of Advanced Computing (CDAC) specialises in Artificial Intelligence System. He has opened new avenues through his extensive research on major R&D projects in Natural Language Processing (NLP), Machine assisted Translation (MT), Information Extraction and Information Retrieval (IE/IR), Speech Technology, Mobile computing, Decision Support System and Simulations. He is a recipient of the prestigious "Computerworld Smithsonian Award Medal" from the Smithsonian Institution, USA for his outstanding work on MANTRA-Machine Assisted Translation Tool which is also a part of "The 1999 Innovation Collection" at National Museum of American History, Washington DC, USA.

Mr. Nisheeth Joshi is a researcher working in the area of Machine Translation. He has been primarily working in design and development of evaluation metrics in Indian Languages. Besides this he is also very actively involved in the development of MT engines for English to Indian Languages. He is one of the experts empanelled with TDIL Programme, Department of Information Technology, Govt. of India, a premier organization which foresees Language Technology Funding and Research in India. He has several publications in various journals and conferences and also serves on the Programme Committees and Editorial Boards of several conferences and journals