International Journal of Academic Research in Business and Social Sciences
2017, Vol. 7, No. 12
ISSN: 2222-6990

# Semantic Relationships Rules Identification for the Construction of Medicinal Herbs Domain Ontology

**Zaharudin Ibrahim[1], Tengku Adil Tengku Izhar[1], Mohd Sazili Shahibi[1], Ahmad Zam Hariro[1], Mohd Ridwan Seman@ Kamarulzaman[1], Mahanem Mat Noor[2], Shahrul Azman Mohd Noah[3]**

[1]Faculty of Information Management, Universiti Teknologi MARA UiTM, Selangor, Malaysia

[2]School of Bioscience and Biotechnology, Faculty Science and Technology, Universiti Kebangsaan Malaysia, 43650 Selangor, Malaysia

[3] Centre for Artificial Intelligent Technology (CAIT), Faculty of Information Science & Technology, Universiti Kebangsaan Malaysia

DOI: 10.6007/IJARBSS/v7-i12/3728 URL: http://dx.doi.org/10.6007/IJARBSS/v7-i12/3728

**Abstract**

The primary goal of ontology development is to share and reuse domain knowledge among people or machines. This study focuses on the approach of extracting semantic relationships from unstructured textual documents related to medicinal herb from websites and proposes a lexical pattern technique to acquire semantic relationships such as synonym, hyponym, and part-of relation. The results seven types of concepts (entities), eight object properties (or semantic relations) and twenty lexico-syntactic patterns have been identified manually, including one from the Hearst hyponym rules. The lexical patterns have linked fifty one terms that have the potential as concepts. Based on this study, it is believed that determining the lexical pattern at an early stage is helpful in selecting relevant term from a wide collection of terms from the corpus. However, the relations and lexico-syntactic patterns or rules have to be verified by domain expert before employing the rules to the wider collection in an attempt to find more possible rules. This study shows that background knowledge about the domain is essential to develop the TBox ontology diagram that serve as backbone of the domain ontology. This diagram is essential as guideline in discovering lexico-syntactic patterns therefore expedite the knowledge extraction process.

**Keyword**: Knowledge Management, Herb, Semantic Web, Natural Language Processing, Biomedical, Knowledge Engineering, Web Documents.

## 1. Introduction

First introduced by Aristotle, ontology has recently become a topic of interest in computer science. Ontology provides a shared understanding of the domain of interest to support communication among human and computer agents; it is typically represented in a machine processable representation language (Haase and Sure, 2004) and is also an explicit formal specification of terms, which represents the intended meaning of concepts, in the

*International Journal of Academic Research in Business and Social Sciences*
*2017, Vol. 7, No. 12*
*ISSN: 2222-6990*

domain and relations among them, and considered as a crucial factor for the success of many knowledge-based applications (Staab et al., 2001). With the overwhelming increase in biomedical literature in digital forms there is a need to extract knowledge from the literature **(**Fuller, et al, 2004). Ontology may also be helpful in fulfilling the need to uncover information present in large and unstructured bodies of text, commonly referred to as non-interactive literatures (Swanson & Smalheiser, 1997 ) i.e., literatures that do not cite each other but which, nevertheless, together present useful new information. Ontology is considered as the backbone of many current applications, such as knowledge-based systems, knowledge management systems and semantic web applications. One of the important tasks in the development of such systems is knowledge acquisition. Conventional approaches to knowledge acquisition are mainly from interviewing domain experts and subsequently modelling and transforming the acquired knowledge into some form of knowledge representation technique. However, a huge amount of knowledge is currently embedded in various academic literatures and has the potential of being exploited for knowledge construction. The main inherent issue is that such knowledge is highly unstructured and difficult to transform into meaningful model. Although a number of automated approach in acquiring such knowledge has been proposed by Alani, et al (2003) and Cimiano, et al (2005) their success have yet to be seen. Such approaches have only been tested on general domain and scientific domains such as the medicinal herbs domain have yet to be explored. While automated approach seems to offer promising solutions, human still play an important role in validating the correctness of the acquired knowledge, particularly in scientific domain. This study, therefore, proposed a semi-automated approach for discovering domain-specific concepts and relationships in an attempt of acquiring domain knowledge of the medicinal herbs domain from web documents. The Hearst's technique (Hearst, 1992) has been employed to extract concept terms from the literature and to discover new patterns through corpus exploration. The technique acquires hyponym relations automatically by identifying a set of frequently used unambiguous lexico-syntactic patterns in the form of regular expressions. The study of lexico-syntactic pattern also was carried out by Moldovan, et al (2000) which discovered domain-specific concepts and relationships in an attempt to extend the WordNet. The pattern-based approach was also applied by Pantel & Pennacchiotti (2006) which proposed a bootstrapping algorithm to detect new patterns in each iteration. There are also studies by Imsombut & Kawtrakul (2007) that proposed methods which are very close to the pattern-based approach of extracting the ontology from item lists, especially in technical documents, and by Zaharudin, et al (2009) which applied a semi-automatic approach in collecting relevant terms extracted from lexico-patterns from medicinal herb documents to be inserted as ontological term.

The main objective of this study is to explore the possibility of identifying semantic relationships (SR) between terms from medicinal herb literature and to represent it in the form of some ontological structure. Since ontology can be represented in many forms, ranging from a simple list of concepts to the complexity of logic structure (Zaharudin, et al, 2009), this study focuses on representing the concepts that was manually selected by expert with semantic links and discover the lexico-syntactic patterns or semantic rules. The rules created from analyzing a set of textual documents will be used to populate medicinal herb domain ontology. The paper

focused on our initial findings based upon current approaches of extracting knowledge from unstructured documents as previously described.

## 2. Materials and Methods

The advances in the biomedical sciences need the development of ontology to help user understand the developments in one's own area of specialization, and also to enable them to quickly learn about the developments in related and unrelated subject areas. Ontology plays an important role in facilitating the formal sharing and re-using of knowledge through the construction of an explicit domain model (Fensel, 2004; Gruber, 1995). However, the manifestation of this role requires the construction of concepts in the particular domain. The meaning of terms and the relationships between entities within that particular domain of knowledge must be well defined. This is to ensure that the knowledge can be interpreted logically before a body of knowledge can be communicated unambiguously across computer-based systems (Catton et al., 2002).

## 3. Rules for Populating Medicinal Herb Ontology

Our approach mainly concerns with identifying entities related to the domain such as ’herb’, ’effect’, ‘usage’ and etc. These entities are related to the TBox ontology which was constructed from consulting domain experts from the Reproductive and Developmental Biology Research Group, Universiti Kebangsaan Malaysia and analysis of reputable literature. Such ontology is illustrated in Fig. 1.

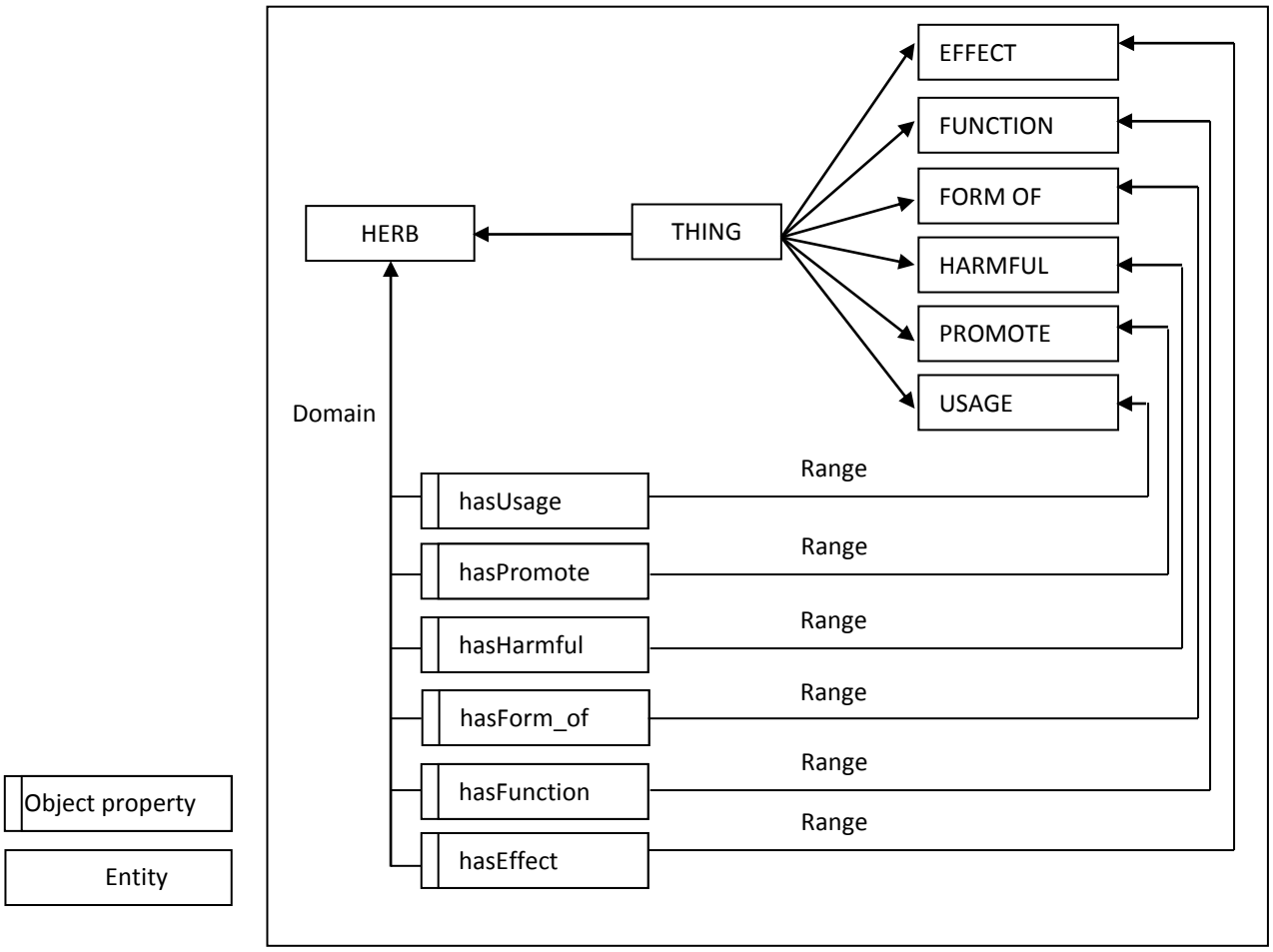


Fig. 1. The TBox ontology for the medicinal herb domain

As can be seen from Fig. 1 there are seven concepts (entities) and eight object properties (or semantic relations) in the constructed ontology. The rules designed for name entities recognition are mainly set of patterns which matched with the given object properties and subsequently inferred to be instances of the corresponding concepts. These nine semantic relations are constructed according to stages illustrated in Fig. 2.

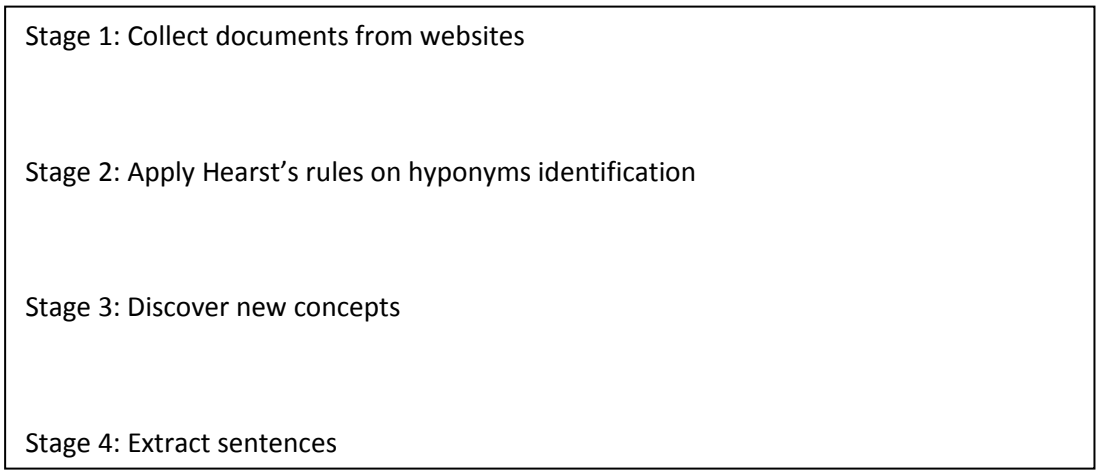


Fig. 2. The stages of the semantic relation identification

For the initial stage, several documents related to Malaysia medicinal herbs have been collected from websites and other resources. We then apply Hearst's rules to identify hyponyms. In this case the collection or corpus was analyzed to extract the semantic relationships enhancement based on the work of [Hearst 1992]. The Hearst rule for detecting hyponym from text includes:

1. NPo . . .such as{NP1,NP2 . . .(and\or)}NP, ,
2. such NP as{NP,} _ {(or [and)}NP
3. NP{,NP} _ {, } or other NP
4. NP{,NP} _ {, } and other NP
5. NP{, } including {NP _ {or\and} NP
6. NP{, } especially {NP,} _ {or] and}NP

**4. Research Approach**

In this study, a semi-automatic approach using NLP tools was used to collect relevant terms in medicinal herb documents to be inserted as ontological term. This approach requires that the researchers be familiar with the terms or concepts in the domain. They should have substantial knowledge regarding the domain before selecting the appropriate terms that can be linked to each other.

*Stage 1: Document Collection and Preparation*

For the initial stage as shown in Fig. 2, several documents related to medicinal herbs have been collected from websites.

*Stage 2: Apply Hearst's rules on hyponyms identification*

The collection or corpus was analysed to extract the semantic relationships enhancement based on the work of Hearst (1992) as described in previous section. In order to find other semantic relation rules or lexico-syntactic patterns, the technique of seed words from Moldovan, et al (2000) was modified as described below.

*Stage 3: Discover new concepts*

In this study a few terms have been selected as seed concepts that are considered important. The focus of this study is on the medicinal herb domain and the seed words selected are *herbs* and the term that mentioned herbs species such as Maca.

*Stage 4: Extract sentences*

The documents retrieved were further processed so that only sentences containing the seed words were retained. Each sentence in this corpus was part-of-speech (POS) tagged and then parsed using Genia Tagger.

Example of the sentence:

Maca is used primarily for enhancing libido and fertility, and treating erectile dysfunction (ED)

Example of the syntactic parser output is:

[NP Maca] [VP is used] [ADVP primarily] [PP for] [VP enhancing] [NP libido and fertility], and [VP treating] [NP erectile dysfunction] ([NP ED])

*Stage 5: Discover lexico-syntactic patterns*

The approach is to search for lexico-syntactic patterns comprising the concepts of interest.

*Stage 6: Extract new concepts*

After discovering lexico-syntactic patterns, the new concepts which are directly related to the seeds were extracted from the sentence. In the example above, the seed word (Maca) was linked with new concepts such as libido and fertility, erectile dysfunction and ED.The analysis processes were simplified into the following stages as shown in Fig. 2.

**5. Results and Discussion**

In this study, several documents from the websites relating to medicinal herbs have been selected for corpus development. Even though Hearst technique intensively applied to linguistic domain but in this the rules was modified to accommodate the purpose of the medicinal herbs domain. The modification involved combination of Hearst techniques and the seed words technique used by Moldovan. The Hearst rule for detecting hyponym from text includes:

(l) *NPo ..... such as {NP1, NP2 . . . . (and /or)} NP,,*
(2) *such NP as {NP ,}* {(or [ and)} NP*

(3) *NP {, NP} * {,}* or *other NP*
(4) *NP {, NP}* {,} and other NP*
(5) *NP* {,} *including {NP** {or / and} NP
(6) *NP* {,} *especially {NP,}* {or] and} NP*

The identification of other relations was heuristically selected if there is a relation between a pair of noun phrase appearing in each of the sentence. The lexical patterns were identified from the relationship between terms (noun phrase) in a sentence. In this study only hyponym has a proper identification approach. Other relations in this list will be investigated further to determine the appropriate relations. This study initially found seven types of concepts (entities) and eight object properties (or semantic relations) in the constructed ontology. The entities are: EFFECT, FORM OF, FUNCTION, HARMFUL, USAGE, PROMOTE and HERB (as shown in Figure 1).

In this study, eight types of semantic relations were found. From these relations, twenty new lexico-syntactic rules were derived for semantic relationships. The details of the rules are given in Table 1.

Table 1. Types of relation terms and rules for semantic relation or lexical pattern

| Relation | Phrase | Rules |
|---|---|---|
| EFFECT | *Nitric oxide , which results from a natural bodily process, **is thought to help** decrease inflammation and increase circulation* | *NP* **is thought to help** {NP1,}* {or/and} { NP}*<br><br>*EFFECT (NP, NP1)*<br>*EFFECT (Nitric oxide, inflammation)* |
| | ***suggests** maca **may enhance** both libido and fertility* | *suggest NP may enhance… {NP1,} {or/and} NP…*<br>EFFECT*( NP,NP1)*<br>EFFECT*(maca, libido)* |
| | Animal research ***suggests that*** maca has estrogen-like effects | suggest that NP has {NP1,}* {or/and} NP…<br>Effect( NP,NP1)<br>Effect (maca, estrogenlike effects) |
| | herbs **that increase** the body's ability | NP that EFFECT<br>EFFECT (NP, NPi)<br>EFFECT (herb, increase the body's ability) |
| | Dong Quai (Tang Kwei) enhances the **utilization of** estrogen | *NP/PN** the **utilization of** * NP {.} {,} {and/or} NP*<br>EFFECT (NP/PN, NP1)<br>EFFECT (Dong Quai, estrogen) |

| | | |
|---|---|---|
| | where Maca **was believed to enhance** fertility, sexual desire, strength and endurance | NP was believed to enhance {NP1,} {or/and} NP<br><br>EFFECT (NP, NP1)<br>EFFECT (Maca,fertility) |
| FORM OF | *These herbs are usually taken in **the form of** a tincture* | *NP/PN * form of {a/an/the} NP1*<br>*FORM OF (NP1, NP/PN)*<br>*FORM OF (tincture, herbs)* |
| FUNCTION | The first recorded **use of** maca **as** an herbal medicine | **use of** NP **as** NP1….<br>FUNCTION (NP, NP1)<br>FUNCTION (maca, herbal medicine) |
| HARMFUL | ***Avoid taking** maca **if** you are pregnant or nursing* | *Avoid taking {NP} if * {NP,}* {or/and} NP*<br>*HARMLFUL (maca, nursing)* |
| | *The following herbs **are not to be used** during pregnancy* | *NP **are not to be used** * NP*<br>*HARMFUL (NP/PN,NP)*<br>*HARMFUL (herb, pregnancy)* |
| | *menopause problems related to low estrogen, **such as** osteoporosis* | *NP such as {NP1}* {(or/and} NP*<br>*Hyponym(NP1,NP)*<br>*hyponym (osteoporosis, low estrogen)* |
| | ***Like other members of** the mustard family, maca* | *Like other members of NP, NP1*<br>*PART OF (NP1, NP)*<br>*PART OF (maca, mustard family)* |
| PROMOTE | *maca **promoted** nitric oxide (NO) production in cartilage* | ***NP promoted NP1…***<br>*PROMOTE **(NP,NP1)***<br>*PROMOTE **(maca, nitric oxide (NO) production)*** |
| USAGE | the herb's most active compounds ***to aid with*** fertility and virility | NP* ***to aid with*** (NP1,)* {or/and} (NP)}<br><br>TREAT (NP, NP1)<br>TREAT (herb's most active compounds, fertility) |
| | *the nourishing herbs **recommended for** fertility **are also** useful during pregnancy* | *NP recommended for * NP1 are also * NP2*<br>*Treat ( NP, NPi)*<br>*Treat (nourishing herb, fertility)* |
| | Maca may also be ***helpful for*** arthritis | NP* **helpful for** {NP1,} {or/and} NP<br>TREAT (NP,NP1)<br>TREAT (Maca, arthritis) |
| | The herb **is** lately being **suggested for** easing PMS and menopause symptoms | NP **is**/are* **suggested*** **for** * {NP,}* {(or/and} NP<br>TREAT(Herb, PMS)<br>TREAT (Herb, menopause symptoms) |

| | | |
|---|---|---|
| | *the nourishing herbs **recommended for** fertility* | *NP **recommended for** * NP1*<br>TREAT*(NP, NPi)*<br>TREAT*(nourishing herb, fertility)* |
| | *A Chinese herb commonly **used to build** the female system, reproductive organ and kidneys.* | *NP/PN* **used to build*** *NP {.} {,} {and/or} NP*<br>TREAT*(PN/NP, NP1)*<br>TREAT*(Chinese herb, female system)* |
| | Maca is used primarily for enhancing libido and fertility, and treating erectile dysfunction (ED) | is **used** primarily **for**<br>PN is used* for * {NP,}* {(or/and} NP<br>TREAT (Maca, libido) |

In this preliminary stage, the ontology populating process has been done manually by mapping all the rules to the documents. As the result about fifty one (51) instances have being extracted as shown in Table 2.

Table 2: List of terms extracted using lexical pattern.

| List of terms | | |
|---|---|---|
| Alkaloid | 18. Herb | 35. Nitric oxide |
| Amino acids | 19. Herbal medicine | 36. Nitric oxide production |
| Arthritis | 20. Herb's most active compounds | 37. Nourishing herb |
| Carbohydrates | 21. Hot flashes | 38. Nursing |
| Chinese herb | 22. Increase the body's ability | 39. Odor |
| Circulation | 23. Inflammation | 40. Osteoporosis |
| Dong Quai | 24. Kidneys | 41. Pregnancy |
| Endurance | 25. Libido | 42. Protein |
| Erectile dysfunction | 26. Low estrogen | 43. Pungent taste |
| Estrogen | 27. Low libido | 44. Reproductive organ |
| Estrogenlike effects | 28. Maca | 45. Saponms |
| Fatty acids | 29. Menopause problems | 46. Sexual desire |
| Female fertility | 30. Menopause symptoms | 47. Strength |
| Female system | 31. Minerals | 48. Tannins |
| Fertility | 32. Mustard family | 49. Tincture |
| Fiber | 33. Natural bodily process | 50. Virility |
| Glucosinolates | 34. Night sweats | 51. Vitamins |

By using the approach proposed in this study all instances that linked by the lexico-syntactic rules have been considered as ontological terms. The knowledge extraction process whereby the instances were discovered is illustrated by ABox diagram as shown in Figure 3.

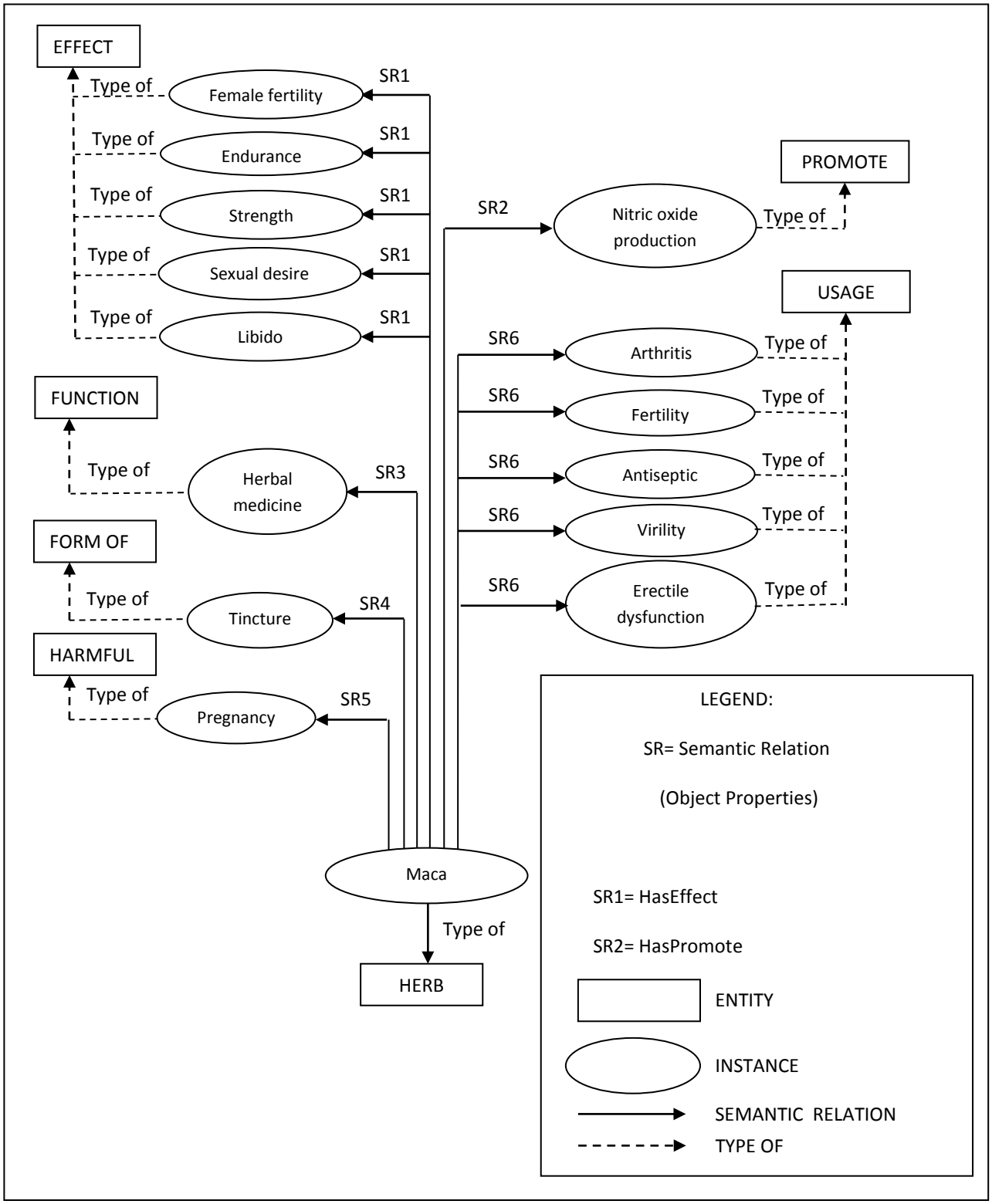


Figure 3. The ABox Ontology Diagram

## 6. Conclusion

In this paper an approach of extracting semantic relation from a specific domain has been presented. This study emphasized on finding the lexical patterns at earlier stage after selecting the relevant terms. This study shows that it is helpful to have background knowledge about the domain prior to finding the lexical pattern. It is hoped that applying all the twenty nine semantic relation rules to other documents in the corpus will reduce the effort to eliminate irrelevant terms which previously has to be done manually. Therefore this approach will minimize the use of domain expert which sometimes is a constraint in the development of the specific domain ontology.

## Acknowledgement

The authors would like to extend our deepest gratitude to the Universiti Teknologi MARA (UiTM) for financing this project under the LESTARI RESEARCH GRANT (Reference 600-IRMI/MYRA 5/3/LESTARI). Our thanks are also dedicated to Research Management Centre (RMC), Universiti Teknologi MARA (UiTM) for facilitating us towards the completion of the project.

## References

Alani, H., S. Kim, D. E. Millard., M. J. Weal.,W. Hall., P. H. Lewis and N. R. Shadbolt, 2003. Automatic Ontology-Based Knowledge Extraction from Web Documents. IEEE. Intelligent Systems. 18(1): 14-21 . http://www2.computer.org/portal/web/csdl/doi/10.1109/MIS.2003.1179189

Catton, C., R. Dalton, C. Wilson, and Shotton (2003). Sabo: A Proposed Standard Animal Behaviour Ontology. Image Bioinformatics Laboratory, Department Of Zoology, University Of Oxford, UK**.** www.bioimage.org/pub/SABO/SABO_cornell_final.pdf.

Cimiano, P., A. Hotho and S. Staab, 2005. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. Journal of Artificial Intelligence Research.24: 305-339. http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume24/cimiano05a.pdf

Fensel, D. (2004). Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Berlin: Springer-Verlag Berlin Heidelbeg New York. ISBN: 3-540-00302-9

**Fuller, S., D. Revere, P. Bugni and G.M.** Martin**,** 2004. A knowledgebase system to enhance scientific discovery: Telemakus. *Biomedical Digital Libraries.* **1:**2. http://www.bio-diglib.com/content/1/1/2

Tagger, G. http://text0.mib.man.ac.uk/software/geniatagger/ index. html

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies. 43(5-6): 907 - 928. ISSN: 1071-5819. http://portal.acm.org/citation.cfm?id=219701

Haase, P and Y. Sure, 2004. State-of-the-Art on Ontology Evolution. Institute AIFB, University of Karlsruhe. http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/SEKT-D3.1.1.1.b.pdf

Hearst, M., 1992. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics. 539-545. http://acl.ldc.upenn.edu/C/C92/C92-2082.pdf

Zaharudin, I., S.A. Noah and M.M. Noor (2009). Knowledge Acquisition from Textual Documents for the Construction of Medicinal herb Ontology Domain. J. Applied Science. 9(4):794-798. http://www.scialert.net/qredirect.php?doi=jas.2009.794.798&linkid=pdf

Imsombut, A and A, Kawtrakul (2007). Automatic building of an ontology on the basis of text corpora in Thai. Journal of Language Resources and Evaluation. 42(2), 137-149. http://www.springerlink.com/content/h24p4p55882704r0/

Moldovan, D., R. Girju, and V. Rus (2000). Domain-specific knowledge acquisition from text. Proceedings of the sixth conference on Applied natural language processing. Seattle, Washington, 268 – 275. http://portal.acm.org/citation.cfm?id=974147.974184

Pantel, P and M. Pennacchiotti, (2006). Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics Sydney, Australia. 113-120. http://portal.acm.org/citation.cfm?id=1220175.1220190

Staab, S., H.-P. Schnurr., R. Studer and Y. Sure, 2001. Knowledge processes and ontologies: IEEE. Intelligent Systems, Special Issue on Knowledge Management. 26-34 . http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00912382

Swanson, D.R. and N.R. Smalheiser, 1997. An interactive system for finding complementary literatures: A stimulus to scientific discovery. Artificial Intelligence. 91(2): 183-203. http://portal.acm.org/citation.cfm?id=251766&dl=