Challenges in Building Domain Ontology For Minority Languages

Panceras Talita
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
94300 Kota Samarahan, Malaysia.
Email: pancerastalita@gmail.com

Alvin W. Yeo
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
94300 Kota Samarahan, Malaysia.
Email: alvin@fit.unimas.my

Narayanan Kulathuramaiyer
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
94300 Kota Samarahan, Malaysia.
Email: nara@fit.unimas.my

***Abstract*—The development of domain ontology is important in building a list of vocabulary whereas the process of sharing and reusing this knowledge management can be accomplished easily. This paper presents the challenges that arise in the ontology development area by focusing on one domain concept. Domain concept here can be transversed from different disciplines, such as agricultural, medicine, human-anatomy, and automotive. The assessments on the challenges vary among numerous ontology projects. The challenges can be influenced by the use of minority languages as the local resources since these languages are resource constrained compared to languages such as English that are rich in resource availability. Apart from that, minority languages tend to have issues concerning different morphological structures and grammatical structures. Numbers of existing ontologies for different disciplines had been produced in English language but little has been done for indigenous languages such as Iban. The main contribution here resides in the ontology development itself, which emphasise on the best means for a beginner to design, develop and deploy the ontology. Research based on the previous work and possible solution is presented in this paper.**

***Keywords*—Domain; Ontology; Natural Language Processing (NLP); Minority Languages; WordNet**

I. INTRODUCTION

In recent years, the use of ontology in the field of Natural Language Processing (NLP) and Artificial Intelligent (AI) has become a necessity in exploiting the information for an efficient and useful management of knowledge. The term 'ontology' can be described as a collection of concepts and relationships among a specific domain application. Besides, ontology is an explicit specification of a conceptualization [1].

Ontologies come in different flavours; from flat lexicons with very few relationships to very expressive ontologies, which attempt to capture every possible aspect of the domain and have broad support for axioms [2]. Most of ontologies share the same content structural, such as classes (concepts), individuals (instances), properties (attributes), relations, and restrictions. The mixtures of the structural components will constitute a knowledge base for the ontology.

On the other hand, the reason for the development of ontologies is to be able to share and reuse the basic understanding between different entities such as human and application. In this paper, we will focus on explain the fundamental challenges face during the ontology's development using textual text form as a resources, for our Iban WordNet. This Iban WordNet (IbaWN) is build based on Iban language which focuses on the agricultural domain. The rest of the paper is organised as follows: Section 2 describes the objective of this paper. Section 3 discusses on the ontology: why choosing domain ontology and explaining in detail about WordNet. Section 4 discusses the main goal and scope of the ontology besides justifying on the techniques and methods which apply in developing the ontology. Section 5 presents the challenges face while building the domain ontology. Section 6 presents the future work of the research and finally, Section 7 is devoted to the conclusions of this work.

II. OBJECTIVE

The objective of this paper is to highlight the challenges met during the phase of constructing the ontology for agricultural domain using Iban as the main language. Iban is one of the divergent Dayak ethnic group in Sarawak, which also comprise other ethnic groups such as Kenyah, Kayan, and Kelabit. This domain ontology later will be used for the development of Iban WordNet (IbaWN). The use of ontology in the WordNet may provide substantial benefit to users in term of:

Describe and represent data in an explicit manner, namely semantic, for a better understand of the knowledge in one particular area;

Permit data interchanging and information sharing between entities, human and application;

To help users in term of learning and understanding the basic structure of the language such as grammar and vocabulary.

III. ONTOLOGY

Ontology is a representation of knowledge, intended to capture the conceptualisation information, like entities, properties, interrelationship, and functions. The growth of ontology is becoming more widespread in many different fields such as information system, natural language processing, information retrieval and extraction, and knowledge management. Numbers of different ontologies exist nowadays; usually built to establish

communication between entities; people and applications for knowledge sharing and understanding. [3] There are five different types of ontologies:

Domain ontologies: Symbolise the knowledge which is associated to a particular type of domain, such as medicine, automobile, and electronics.

Generic or *Common Sense ontologies*: Exploited to capture the general knowledge about the world, such as time, space, state, and events.

Metadata ontologies: Used to describe the content of on-line information sources.

Representational ontologies: Not "obligated" to any particular domain, as it provides the representational entities without starting of what should be represented.

Method ontologies: Describing specific term for a particular task.

Table I, shows the example of existing and establish ontology for differences types of ontologies.

TABLE I.
ONTOLOGY SAMPLE BASED ON ONTOLOGIES TYPES

Ontology Type	Example
Domain Ontologies	Gene Ontology Url:http://www.geneontology.org/
Generic Ontologies	GDM Ontology Url: http://www.egovpt.org/fg/
Metadata Ontologies	Dublin Core Url: http://dublincore.org/
Representational Ontologies	Frame Ontology Url: http://www-ksl.stanford.edu/
Method Ontologies	Method Ontology http://ksi.cpsc.ucalgary.ca/

Generally, the level of generality and the level of detail are drawn on to classify and distinguish the distinction amount of knowledge specified in the ontology. For the remainder of this paper, we will clarify the methods and challenges in the development process of our domain ontologies.

A. Why Domain Ontology ?

People tend to be confused between the term 'Domain Ontology' and 'Upper Ontology'. *Upper ontology* refers as the top-level ontology. The upper ontology is an ontology which describes the general and common concepts that are similar across all the domain ontologies. The most basic function of upper ontology is that it supports an extremely broad scope of semantic interoperability which is useful for semantic foundation on semantic web. This semantic foundation here covers: searching, communication and interoperation. On the other hand, *domain ontology* is characterised as a more specific domain that symbolises a more particular meaning towards the terms. This domain ontology describes the vocabulary related to the specific field or an application.

In this case, the ontology described is the agricultural domain for Iban language. The objective of this ontology is to produce a vocabulary list on agricultural domain in Iban which in due course assist in the process of preserving the language itself.

B. WordNet

Today, there exist at least 40 WordNets in different languages. The initial WordNet, known as Princeton WordNet (PWN) was originally created for the English language and nowadays, expanding to several languages worldwide due to its utility. Similar projects cover languages such as for Korean, Japanese, Arabic, and Spanish.

WordNet is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory [4]. WordNet is a collection of large linguistic data containing individual senses (words) connected to one another by relations (semantic). In WordNet, a word is tied to a definition or gloss and can have one or more part-of-speech (POS) categories such as noun, verb, adjective, or adverb. Furthermore, each of the categories organise the words according to concepts or word meaning based on the semantic relationship amongst words. A popular tradition of studying semantic representation has been driven by the assumption that word meaning can be learned from the linguistic environment [5]. These are several examples of the relationships:

- *Synonymy*: Have similarity in meaning of the words, which is used to build concepts represented by a set of words.
  Example: 'Black' and 'Charcoal' are synonyms

- *Antonymy*: Have an opposite in meaning of words which are mainly used for organising adjectives and adverbs.
  Example: 'Black' and 'White' are antonyms

- *Hypernymy*: Kind-of relationship refer to a hierarchical relationship between words.
  Example: 'Tree' is a hypernym of 'Oak'

- *Meronymy*: Part-of relationship between concepts.
  Example: 'Wheel' is a meronym of 'Car'

By inferring these relationships, a model of semantic network structure is establish as from what we can see for English language in PWN and others as well. Figure I shown the model structure for the ontology:

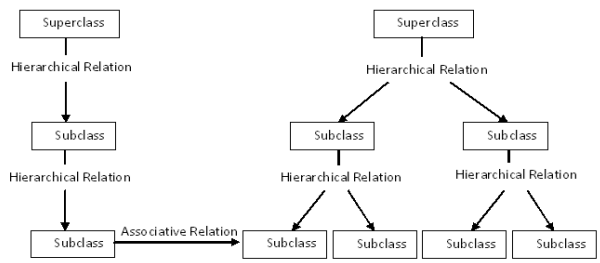

Figure I. Ontology Structure Model

IV. BUILDING ONTOLOGY FOR THE AGRICULTURAL DOMAIN

This section describes the goal and scope, besides the implementation procedure used in building the domain ontology for the agricultural domain. This ontology is useful in designing the WordNet's databases and knowledge bases.

A. Goal and Scope

The goal and scope are interrelated to one another. The scope of the ontology should be sufficiently well defined to ensure that the depth and detail are covered. This will put a boundary on the ontology whereby, only the specified information addresses in the goal is included. Here, only a significant amount of data and concept is used to be analyzed, namely to reduce the impact and complexity towards the agricultural semantic concepts.

This ontology will focus on creating the agricultural domain with concepts related and limited to; agriculture and fishery industry. Other concepts related to the agricultural domain such as forestry will not be included. As a result, a general concept on the agriculture and fishery are specific based on inspection on the resources collected covering information, including human action and behaviors. This domain is chosen as the resources on the subject of the agriculture assessment and aspect can be acquire and gather without required much effort.

B. Implementation

Researchers who deal with the challenge of building ontologies from text can choose from an essentially considerable array of methodologies used in computational linguistics and natural language processing [6]. This ontology development deals with the process of modeling the world with a shareable knowledge and information structures. Several ontology building techniques, as automatic or semi-automatic has been the subject in many research fields where most of the ontologies are currently build via manual effort. Beforehand, several different development methodologies were studied [7].

Methontology [8]: This methontology framework provides automated support for ontology development based on IEEE standard for software development.

Lexicon-Based Ontology Development Method [9]: This ontology construction process concentrates on establishing the requirements elicitation strategy focusing on application languages.

Toronto Virtual Enterprise Method [10]: This method build based on previous experience by manipulating the similar scenarios to describe the problems and examples that were not addressed in any existing ontologies.

Ontology Development 101 [11]: This ontology development method is proposed as a set of guidance for the users to create their first ontology.

From the above study, we have identified that ontology development method can be exactly divided into two main methods. These methods are experienced based methodologies (Toronto Virtual Enterprise Method) and evolution prototypes model (Ontology Development 101 and Methontology).

In this research, the ontology development technique employ is the user-centred methodology [12] based on the use of Machine Learning and Natural Language Processing. The detail of the development process is basically describes as follow:

Step 1 (Planning): Planning the preliminary design of the ontology by sketching up the ontology to obtain overall overview of the structure and relationship. A set of textual text form on the relevant domain is process to produce a corpus.

Step 2 (Discovery): Applying ontology in order to identified semantic relationship. Domain expert and lexicographer is use to perform verification and validation on the instances.

Step 3 (Enrichment): A pattern will be generated based on the validated relationship. This pattern helps to enrich our ontology by finding other existing instances with the same relations.

Step 4 (Populate): The discovery of new instances will be repeated until best suit ontology is achieved. This automate process will help build up the ontology without consuming much time and cost.

A tool, GATE [13] is used to simplify this process. This tool help in developing and exploiting the knowledge resides in the textual form. Besides, GATE is capable in maintaining and evolving the ontologies and metadata over time. Table II, shown the basic features of GATE:

TABLE II.
AVAILABLE FEATURES IN GATE

Features	Description
Tokeniser	Split the text into simple tokens such as numbers, punctuation and words of different types.
Sentence Splitter	Help segment the text into sentences.
Gazetteer	Identify entity names in the text referring to the lists. The lists can represent a set of names, such as names of the countries and cities.
Part-of-Speech tagger	Generates part-of-speech tag as an annotation on each word or symbol.

The advantages of using GATE as the text processing tool is that these tools can easily be integrated, customised, and support different kind of languages and document formats.

V. CHALLENGES IN BUILDING DOMAIN ONTOLOGY

A. Lack of Well Defined Semantic

One of the significant challenges encountered during the process of developing the domain ontology usually concerns the lack of well defined semantics. WordNet is a well defined lexical database which structures the

information in a semantic manner. In this research, we have to deal with the disorganized textual form contain as our resource, such as books (e.g.: story books and dictionaries), magazines and newspapers which are not adequate with any semantic representation. Without semantic representation, the process of extracting and exploiting the resources will became harder, thus would lead to other problems such as time consuming.

The resources collected are mainly in Iban language. Currently, not much resource is available for Iban language that caused the limitation of the scope to one particular domain, namely agriculture.

Hence, there is a great need for this available textual form to be exploited for information accession and extraction. This process is vital whereby; the ontology backbone is built by the process of building and shaping the database itself.

Semantic are necessary to enable the entities: applications or human to understand and determine which lexical to be stored as ontology. With the issues related to semantic representation, the quality and accuracy of the ontology will be a real concern.

B. Ambiguity between the Association Terms

This problem happens when different people have different associations with one particular term. In our case, Iban language and culture are different according to the place they settle in. Iban can be divided into different branches according to region, such as Iban Sebuyau, Iban Saribas, and Iban Balaus. Iban Sebuyau for example, lives around the Lundu and Samarahan area in Sarawak whereby Iban Saribas come from the Betong and Saratok area also in Sarawak. Table III, shown the term variation based on different Iban dialects:

TABLE III.
TERM VARIATION BASED ON DIFFERERENT IBAN DIALECTS

English	Iban (Samarahan)	Iban (Saribas)	Iban (Kanowit)
Kitchen	Dapuh	Dapor	Dapor
Rice	Behas	Beras	Berau

This has been a challenge to determine the correct synsets and avoid any ambiguity in representing the accurate term in Iban language. This term is necessities to help match the definition associated with it.

A standard, as in our case standard dialect is important if the particular language itself possesses different numbers of dialect. Furthermore, ontology mapping can be done by mapping the taxanomy or thesaurus between two different ontologies for two different languages [14], describes semantic differences among thesauri that affect the mapping process.

Besides being a great help in building our domain ontology, this standard without doubt can standardise the language itself for further preservation and revitalisation.

C. Lexicography

The lexicography is one of the common problem exists in any domain ontology development. As we know, the way the people speak their local language and practice the culture is very much different depending on the location and custom. Each community in these different places possesses its own dialect. The issues such as the different dialects will affect the quality and accuracy of the ontology.

Here, selected lexicographers which are the members of the ethnic group, constantly face issues of misunderstanding. The reason occurs because the difference dialects. One lexicographer could be members of Iban community living in the area of Sri Aman, Sarawak and the other could be from the area of Kapit, Sarawak. A process of negotiation may be required to come to an agreement on which terms to use for the different ethnic groups.

Hence, the translation and enriching process of the concepts will be a problem. The translation could vary for one single term and also the possibility of unnecessary concept could exist. This leads to the drawback factors such as inconsistency, inaccuracy and inefficiency which reduce the level of ontology performance. Table IV, shown the differences term based on different native speaker:

TABLE IV.
DIFFERENCES TERM BASED ON DIFFERENT NATIVE SPEAKER

Malay Speaker	Iban Speaker	Justification
/lidah/	/dilah/	Same sound and meaning
/jari/	/jari/	Same sound different meaning

This matter can be overcome using the dictionary. The dictionary can be categorized into many types: monolingual, bilingual, and general. The types of dictionary which can address this problem is the terminological dictionaries where its apply concept relationship. There no terminological dictionary available for Iban language. Manual alignment using monolingual and bilingual dictionary will be used in our research.

D. Time Constraint and Cost Expensive

Although ontology has play an important role in most of the Natural Language Processing (NLP) and Artificial Intelligent (AI) area, the building process has become the stumbling block when considering the time and cost factor. The ontology development cost is subjected to the aspect of ontology maintenance, reuse and quality.

- *Ontology Maintenance*: Can be defined as the procedure of modifying which include the process of inserting and deleting the ontological primitives and also due to the remodeling aspect.
- *Ontology Reuse*: Involve the process of discovery and reuse the current existing ontologies to generate new ontology.
- *Ontology Quality*: Involve with the aspect of richness, correctness, and coverage.

Besides, the ontology development time is subjected to the level of complexity and building method used in developing the ontology.

- *Ontology Complexity*: Analysing and control of the level of ontologies's complexity is crucial in

ensuring the ontology is useable for other type of application.

- *Ontology Building Method*: Involve the process of selecting the best method, either a semi-automatic or manual method.

There is no simple solution to help facilitate the process of developing and enriching the ontology. Most of these problems had become the major drawback to the researches in their effort to overcome the time consuming and cost expensive problems. Several methods and tools are available for use in order to minimize and reducing the impact of this problem.

FUTURE WORK

The outcome of this research will provide us with a software framework for building WordNet in Iban language. This WordNet will contain the synsets related to the agricultural domain. This agricultural domain will act as an ontology whereby related structural information is available for people to use as efficiently and accurately. Here on, we will then adopt this software framework to build WordNet for other languages in Sarawak, such as Melanau, and Kayan. This will absolutely assist the community, especially the local people who are keen to preserve, maintain and revitalise their local language.

CONCLUSION

In this paper, we have shown that building the domain ontologies is not an easy process. Ontologies developers need to have background knowledge in identifying the suitable technique and method in designing the ontology. This can help in minimizing the affects such as problems and constraint that might question the ontologies quality.

Here, we emphasis in explaining the challenges which take place during the establishment of the agricultural ontology. Most of the challenges are affected by the non-human factor, for example the characteristic of the language itself, in term of the dialect and linguistic uniqueness. Although some challenges are caused due to human factor, such as lack of knowledge and expertise as a domain expert and lexicographer.

The information and knowledge we gather here can be used by others, especially the community that involved in ontology construction project. Hence, the contribution that this paper offers is the basic overview of what the problems and possible solutions each of us can use as the guidance towards the development of the domain ontologies for indigenous languages.

REFERENCES

[1] Gruber, T. 1995. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human and Computer Studies, 43(5/6): 907-928.

[2] Jos de Bruijn. Using ontologies - Enabling Knowledge Sharing and Reuse on the Semantic Web. Technical Report DERI-2003-10-29, DERI, 2003.

[3] D. Fensel. Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Heidelberg, Germany, 2001.

[4] Christiane Fellbaum (ed.): WordNet: An Electronic Lexical Database, MIT Press, 1999.

[5] Yansong Feng and Mirella Lapata. 2010. Visual Information in Semantic Representation. In: Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 91-99. Los Angeles, CA.

[6] Buitelaar P, Cimiano P, Magnini B. Ontology Learning From Text: Methods, Evaluation and Applications (2005) IOS Press.

[7] Wache H., Vögele T., Visser U., Stuckenschmidt H., Schuster G., Neumann H., Hübner S. (2001) Ontology-Based Integration of Information –A Survey of Existing Approaches. Proc. IJCAI-01 Workshop: Ontologies and Information Sharing, Seattle, WA, 108-117.

[8] Fernandez Lopez M, Gomez Perez A, Juristo N (1997) METHONTOLOGY: From Ontological Art Towards Ontological Engineering. Spring Symposium on Ontological Engineering of AAAI. Stanford University, California, pp 33–40.

[9] Breitman, K.; Leite, J.C.S. (2003) Ontology as a requirement engineering product. In: Proceedings of the Eleventh IEEE International Requirements Engineering Conference. 8-12 Sept. 2003, Monterey Bay, California, USA, pp. 309–319.

[10] Gruninger, M.; Fox, M. (1995) Methodology for the design and evaluation of ontologies. In: Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, in IJCAI-95, Canada.

[11] Noy, N.; McGuiness, D. (2001) Ontology Development 101 – A guide to creating your first ontology. KSL Technical Report, Stanford University, Stanford, CA, USA.

[12] C. Brewster, F. Ciravegna, and Y. Wilks, 'User centred ontology learning for knowledge management', in Natural Language Processing and Information Systems, 6th International Conference on Applications of Natural Language to Information Systems (NLDB 2002).

[13] GATE. Available. Retrieved on 14 March 2010 from: http://gate.ac.uk/

[14] Doerr, M. 2001. Semantic Problems of Thesaurus Merging. Journal of Digital Information. 1(8).