A REVIEW ON LEARNING TAXONOMIES FROM MALAY TEXT CORPORA

Mohd Zakree Ahmad Nazri[1], Siti Mariyam Shamsudin[2], Azuraliza Abu Bakar[3]

[1&2]Department of Graphic & Multimedia
Faculty of Computer Science &
Information Technology, Universiti
Teknologi Malaysia
[1]mzan@ftsm.ukm.my
[2]mariyam@fsksm.utm.my

[3]Department of System Science &
Management, Faculty of Information
Science & Technology, Universiti
Kebangsaan Malaysia
aab@ftsm.ukm.my

Abstract: Taxonomy is a science of classifying living things. In the 21st century, taxonomy is also known as a form of business intelligence, used to integrate information, reduce semantic heterogeneity, describe emergent communities and interest groups, facilitate the communication between information systems. However, in building a taxonomy, knowledge acquisition is the bottleneck that . Ontology engineers also need guidelines about the effectiveness, efficiency and trade-offs of different methods in order to decide which techniques to apply in which settings. But there are no comparative work systematically analyzing different techniques and algorithms on learning concept hierarchies from a Malay text. In this paper we review the state of the arts in taxonomy learning and address the lack of work in the field of concept hierarchy induction from Malay text. We also define an evaluation methodology to systematically comparing different approaches. In our further works section, we proposed an experimental approach to study various approaches and methods to automatically acquire concept hierarchies from Malay texts.

Key words: Ontology Learning, Artificial Intelligence, Natural Language Processing, Formal Concept Analysis, Taxonomy

1.0 INTRODUCTION

Domain taxonomy is the first step towards effective classification and retrieval, concept sharing, interoperability among machines, web communities, corporate enterprises and interest groups [1]. Taxonomies have emerged from specific fields like zoology, biology and library science into Information Retrieval, Text Mining, Natural Language Processing. Taxonomy which evolved from life sciences is now considered the backbone of an organization's information architecture. For instance, taxonomy is an integral part of a content management system.

The notion of a new generation of WWW called Semantic Web, has demonstrated a paradigm shift in communications by changing the WWW from mere static information displays into machine readable content. Realizing the potential of Semantic Web, various ontologies in English language has been developed for many areas such as UNSPC and Rosetta Net ontology for e-commerce, Gaeln and UMLS ontology for medical, Engmath ontology for engineering, Gene Ontology for bioscience and Semantic Bible project for theology. However, contrary to the current trend in ontologies, we find no evidence for the existence of ontology in Malay language. For example, in the context of Islamic domain, the most structured representation of al-Qur'an (English Translated) in Web can only be found in the XML version by Jon Bosak (1998).

Although one of the main purposes of taxonomies is to reduce the effort during the knowledge acquisition process, acquiring knowledge for building a taxonomy from scratch is time consuming and expensive [2]. Even though ontology development tools such as Protégé-2000, Ontosaurus and OilEd have matured over the last decade, it still remains a tedious, cumbersome task which can easily result in a knowledge acquisition bottleneck [3], i.e. the difficulty to actually model the domain in question. Besides, neither of the tools is designed with specific intention of handling and representing knowledge in Malay. Furthermore, after an extensive research within the existing taxonomy and ontology learning literature, we are not aware of any effort towards learning concept hierarchies (taxonomy) from a Malay text corpora using any available technique.

This paper is organized as following. Section 2.0 presents a summary of literature review relating to the research to be pursued. Section 3.0 will be discussing the strength and weaknesses of current approach to solve the problem and other issue that need to be explored/study. Section 4.0 will propose the research approach and methodology in solving the problem. We conclude this paper with section 5.0.

2.0 LITERATURE REVIEW

Wikipedia defined taxonomy as the science of classifying living things [4]. Taxonomy is commonly in hierarchical structure which representing relationships between concepts within a defined scope and context. In the Methontology framework, building glossary of terms and then taxonomy is considered the first and second step respectively towards creating a *formal ontology* of a domain. [5] supported this framework by expressing an ontology as:

Ontology =<taxonomy, inference rules> , and

Taxonomy = <{classes},{relations}>

On the contrary, [5] stated that there are people who consider taxonomies a full ontology. But according to [2], the ontology engineering community prefers to categorize taxonomy as 'light ontology' which means an ontology without axiom and rules (constraints). The word Ontology has a long history in philosophy which refers to the systematic explanation of being. From an IT industry perspective, the word ontology or 'applied ontology' is used to describe the linguistic specifications needed to help machines effectively share information and knowledge [5]. The most quoted definition of ontology is '*ontology is a formal specification of a conceptualization* by Gruber (1993).

This section is partially based on material already published in [1-3, 6-8]. Learned from the material cited before, we divided our research into a two stage process: *term extraction* and *concept formation and hierarchy induction.*

In the first stage, we need to process the corpus and turn it into a form that learning methods that depends on syntax features such as clustering, Formal Concept Analysis or classification can use. The literature provides many examples of Malay term extraction methods that could be used as a first step in taxonomy learning from text. Most of these are based on information retrieval methods [9-12] but many are inspired by terminology and NLP research. Research in computational linguistic for processing English corpus are a mature field. In [3] for example, they used existing tools such as TreeTagger [13] to build a parse tree for each sentence and parsed using LoPar [14] for Part-of-Speech (POS) tagging. This stage is very crucial as the effectiveness rate of learning depends on the quality of outputs of the NLP process done during this stage. Term extraction implies more or less advanced levels of linguistic processing, i.e., phrase analysis to identify complex noun phrases that may express terms and dependency structure analysis to identify their internal structure. As such parsers for Malay is not available like LoPar, much of the research on this stage in taxonomy learning has remained rather restricted. The common approach is mostly to run a POS tagger over the corpus and then to identify possible terms by manually constructing ad-hoc patterns. In order to identify only relevant term candidates, a statistical processing step may be included that compares the distribution of candidates between corpora using for example a χ^2 test or similar.

However, Malay POS tager that need to be built must be done accordingly to Malay

grammatical rule. According to the research done by [15], there are three types of Malay language grammar. First, sentence grammar (SC) developed by [16] and [17], second, partial discourse grammar [18] and third, 'pola' grammar. The SC is based on the models of transformation-generative grammar and the relational grammar of English [2]. The developed SC were inherently from the phrase structure grammar (PSgrammar) developed by N. Chomsky in 1957 [11, 16]. The PS-grammar which governs the formation of PSrules results in regular, context-free grammar (CFG), context-sensitive, and unrestricted PS-grammar. To verify the syntax, a CFG for Malay language was developed by [19]. The CFG describes the structure of basic sentence in Malay language which is either a combination of NP+NP, NP+VP, NP+PP, or NP+AP. For example, the parse trees in Fig. 1 show the derivation for the basic sentences "*Dia menulis aturcara* " or "He writes (computer) program".

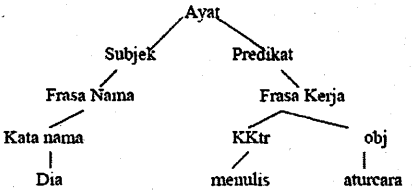

Fig. 1 A Derivation of Sentence

(Fig source : [11])

There are numbers of research used CFG in their conflation method to increase Malay-based IR effectiveness like [9] [10] [20] and [21]. Stemming algorithm is used to conflate morphological variants. For English as example, Porter's Algorithms stems only suffixes [12]. Compared to Malay language, we have more than just the suffixes. One of the earliest work on Malay stemmers for Information Retrieval (IR) was the Othman's Algorithm [20] which has 121 rules for prefixes, suffixes and infixes; Fatimah's Algorithm [9] which has 561 rules and a root words dictionary and Pouzi's Algorithm [10] which integrate WordNet to Fatimah's algorithm to find semantic. Juzaudin's Algorithm [11] which is based on 'pola' grammar is the latest development on Malay's syntactic analysis algorithm.

The second stage, concept formation and hierarchy induction (taxonomy), is a relatively mature field. We have categorized common methods and approaches used in this stage into four areas:

i) Machine Readable Dictionaries : Dictionaries and even textbooks contain explicit knowledge in form of definitions such as *a cat is a mammal.* In many cases, the head of the

first noun phrase (NP) appearing in the dictionary definition is the fact hypernym. For example shown by [6] : *"spring the season between winter and summer and in which leaves and flower appear"*. Some researchers such as [22, 23] have exploited such regular patterns to discover taxonomic or part-of relations in texts. The core idea is to exploit the regularity of dictionary entries to find a suitable hypernymy and also hyponymy relations for the defined word. This method suggests that one can extract frame-based or feature-like structures from dictionaries containing wealth of semantic relations linking the different words together [6].

ii) Latent-Semantic Paten: *Hearst* [24] suggested the application of LSP to the task of automatically learning hyponym relations from corpora. In particular, she defined a collection of patterns indicating hyponymy relations. An example of such a pattern used by him is the following:

such NP_0 as NP_1,..,NP_{n-1}(or|and)other NP_n

If such a pattern is matched in a text, according to Hearst we could derive that for all $0 < i < n$ hyponym(lemma(head(NP_i)), lemma(head(NP_0))), i.e. lemma(head(NP_i)) is a hyponym of lemma(head(NP_0)), where lemma(head(NP)) denotes the lemma of the nominal heads of NP. There are six patterns used by Hearst to find taxonomy relations from corpora. The most recent application of Hearst known to our knowledge is by [25] and [26].

iii) Co-occurance Analysis: Some research has hypothesized that the fact that the occurrence of some word implies the occurrence of some other word in the same sentence, paragraph or document hints at a potential directed relation between both words [6]. Directed means for example a *sub-topic*, *is-a* or *part-of* relation. This notion is related to the one of a collocation. [6] say that two words form a collocation if they occur together in a paragraph, sentence, document or next to each other more than predicted by chance. Sanderson. and Croft, 1999] present a document-based definition of subsumption according to which a certain term t_1 is more special than a term t_2 if t_2 also appears in all the documents in which t_1 appears. On the basis of this document-based definition of subsumption, they automatically induce a hierarchy between nouns from a document collection.

iv) Distributional Similarities: The majority of methods propose a combination of statistical and natural language processing techniques. This is an alternative in deriving knowledge from texts by analyzing how certain terms are used than to look for their explicit definition. The basic intuition behind this representation is the so called *distributional hypothesis* by *[27]*, which states that terms with similar distributional patterns tend to have the same meaning Basically, this approach can be grouped into three classes: the linguistic approach, similarity-

based methods and the set-theoretical. Both later methods adopt a vector-space model and represent a word or term as a vector containing features or attributes derived from a certain corpus.

Linguistic Approach: This approach is named linguistic in the sense that they directly exploit linguistic analysis to derive taxonomic relations. They differ from the clustering approaches described later in that linguistic analysis is not merely exploited for feature extraction, but in a more direct manner. However, this approach seems to rely on other sources to learn relations. [28] for example relies on WordNet to derive the meaning of complex term while [29] present an approach to automatically learn a taxonomy from the search engines such as Google, Yahoo, Clasty and AlltheWeb by given a certain seed of word.

Similarity-based method: Also quoted as clustering method by [6], it is characterized by the use of a similarity or distance measure in order to compute the pair wise similarity or distance between vectors corresponding to two words or terms in order to decide if they can be clustered or not. Some prominent examples for this type of method have been developed by [3, 7]. Reinberger et al [30] has tested and compared results between hard & soft clustering and also studies different kind of weighting measures to weight the significance of the verb-object and verb-subject relation.

Set-theoretical approaches: This approach is based on the strict set-theoretical point of view in the mathematical theory of Formal Concept Analysis (FCA) for the automatic acquisition of lexical knowledge from unstructured data. The main innovation is that the mathematical theory of Formal Concept Analysis is used to extract the monotonic inheritance relations which are inherently given by the data [31]. It has been applied by [3, 31] in learning concept hierarchy.

3.0 DISCUSSION

We have seen that there is a variety and wide scope o techniques which has been applied to the problem of taxonomy learning from English and other European languages textual data. Unlike English, French and Slovene which fall under the Indo-European language family and Arabic which fall under the Semitic language family, Malay language or known as Bahasa Malaysia (BM) falls under the Austronesian or Malayo-Polynesian. Regardless of being described as "the world's easiest language" [32] the morphological structure of Malay words is more complex than English and Slovene [9]. For example, automatic removal of suffixes of an English word to obtain the words' stems is sufficient as variant word forms are created by

adding suffixes to a basic stem. However, for the Malay, stripping the suffixes alone to obtain the root words will not be a sufficient method as the usage of Malay affixes is more complicated than English and Slovene [9]. For BM, obtaining the correct root for each word is regarded as crucial in extracting concepts.

Currently to our knowledge, there are many Malay stemming algorithm such as [9, 10]. But these approaches are meant for Information Retrieval System (IRS). These Malay stemmers are not tested for taxonomy learning. However, in those researches, in order to derive a parse tree for a syntactic process, the CFG was found to be complicated due to many ambiguities for POS. A parser in this stage is important in order to derive a grammar with CFG. It shows the validity of phrases exists in the sentence so that we can POS tag correctly. But, the main problem in developing an effective parser is the ambiguity problem. For example, the word "gemar" can either be a "verb" or an "auxilliary" [11]. In this case, it is called a word ambiguity where a word is ambiguous if they hold more than one part-of-speech (POS). The invalid production will cause a wrong process or an ill-grammar problem. Ill-grammar problems in a POS tagger will results ineffective learning method.

Conventional IRS applies algorithms that can only approximate the meaning of document contents through keywords approach using vector space model. Keywords may be un-stemmed or stemmed. Word stemming is a process in morphological analysis under natural language processing, before syntactic and semantic analysis. [10] and [21] for instance have incorporated stemming in their experimental systems in order to measure retrieval effectiveness. Even though researchers in ontology learning have been using the same measurement like tf-idf for measurement, it is thus unclear how effective they could be used to learn taxonomy from a BM text.

As our intention is to semi-automatically learning concepts and its hierarchy from a Malay corpora i.e a translated al-Qur'an, thesis abstract, etc., the methods based on MRDs and any other approaches that reuse other ontologies such as WordNet or EuroWordNet are less interesting for our purposes.

[6] sees two main drawbacks in using a dictionary-based approach to ontology learning. First, the acquired knowledge heavily depends on the intrinsic idiosyncrasies related to the writing of the entry. Second, in our problem we are mostly interested in acquiring domain-specific knowledge as we have limited numbers of expertise to validate the taxonomy relations acquired. But dictionaries are generally domain independent sources. It is thus unclear how effective dictionary could be used to learn a domain specific taxonomy.

Referring to Table 1, there are numbers of approaches and methods that reuse WordNet to extract relations to a given word or even compared their results using their method to WordNet like [24] and [10]. [10] used an electronic Malay-English dictionary to translate a Malay word into English and submit the word to WordNet to identify the word's hypernym, hyponym and meronym to understand the semantic of the word. We have executed a small scale experiment in finding the semantic of the word 'Nabi' and 'Rasul'. Both are an important concept in Islamic theology which means prophet. Both 'Nabi' and 'Rasul' are translated as prophet which is true but 'Rasul' is a different concept as they inherits all attributes of a 'Nabi' but plus a number of special divinely attributes which other prophet didn't have. Moreover, when the word 'prophet' is submitted to WordNet, the Web based WordNet displayed a foreign concept of prophet to Muslims for example prophetess (a woman prophet) is a hyponym of prophet (which is unfeasible to Islamic teaching). Another hyponym for a prophet is a sibyl (a concept from ancient Rome which means a woman who is considered an oracle). Thus, we concluded that reusing another ontology which is based on western culture and believe for learning taxonomies from Malay or Islamic oriented text is impractical and incorrect.

Referring to Hearst [24] method which we have determined is not a suitable candidate as a solution to our problem, [6] also states that Hearst approach has another disadvantages. The Lexico-Syntactic patterns is however appear rarely and most of the words related through an ***is-a*** relation do not appear in Hearst-style patterns. Thus, according to [6] one needs to process large corpora to find enough of these patterns.

On co-occurrence analysis paradigm, [6] suggested that some research asserts that this depends on the context we're processing. Which means processing (learning) by bigrams, sentences, paragraph or even whole documents may producing different types of relations. [6] claimed that this hypothesis has not been empirically analyzed and therefore we deduce from this claim that there is no empirical research done in the context of Malay text.

Another popular method in taxonomy learning is statistical and machine learning methods. These methods are mostly based on the analysis and comparison of contextual features of terms, extracted from their occurrences in texts (see [7] for a comparison of different vector-based hierarchical clustering algorithms). However, it has its own disadvantages. Taxonomies obtained through this approach, for [1] are very hard to evaluate by a human judge, since kind-of relations are learned on the basis of statistical measures, prone to noise and idiosyncratic data.

[7] have systematically compared different hierarchical clustering approached with respet to effectiveness, speed and traceability by an ontology engineer. An important facts we can learn from these literatures are set-theorethical approaches as FCA can compete and outperform similarity-based approaches in terms of quality of the produced hierarchies, speed and traceability. The beeter quality according to [6] is mainly due to higher recall of the FCA-based approach. Though FCA is exponential in the size of formal context, [6] have shown in his research where the contexts are typically sparsely populated, FCA performs quite efficiently compared to an agglomerative clustering algorithm. Our corpus especially a Malay translated al-Qur'an is believed to be sparsely populated and we believe that FCA is the most appropriate approach to be the first approach to be tested on a Malay text. However, the advantages of unsupervised techniques are also known to us. Applying unsupervised techniques to the generation of concept hierarchies may generate different. problems as [6] already warned us in his literature, that is, spurious similarities and lack of intensional descriptions.

Finally, if we refer to Table 1, performance is often measured with reference to the judgment of human evaluators between 2 and 3 experts [1], usually the authors themselves, where systems are more objectively compared against the same, professionally developed, test set. This information is very important to support our research methodology as we don't have any previous research results, tested and proven data set and text corpora that we can use to benchmark our approach. We have requested two Islamic scholars from Universiti Kebangsaan Malaysia and Universiti Sains Islam Malaysia to be the experts and evaluators as part of our methodology.

4.0 FURTHER WORK

Different approaches and methods have been explored in the literature. We conclude that to our knowledge there is no work concerning the task of automatically learning concept hierarchies beginning from a Malay text, beginning from the term extraction until formal representation of taxonomy. Therefore, there is no comparative work concerning this task. Thus, we proposed an experimental approach in finding the most effective method of learning concept hierarchies from a Malay corpus – specifically in the domain of Islamic Jurisprudence from a Malay translated al-Qur'an. This effort is necessary to break new ground of research in taxonomy learning exclusively from Malay text which will lead a way to far complex process of ontology learning. This endeavor is vital in advent of Semantic Web and sideways with [7] who asserts that ontology engineers need guidelines about the

effectiveness, efficiency and trade-offs of different methods in order to decide which techniques to apply in which settings.

The first component of this research is the text corpora. The first document collection was acquired from [10]. It contains a Malay translated verses of the holy Qur'an. We have stemmed the corpus and there are 205,116 term stemmed from the corpus. Other document collection that we will process and used in this research is a collection of thesis abstract acquired from [11] and Dewan Bahasa and Pustaka articles obtained from [10]. Therefore, we have varieties of writing style which reflects the different domain in Malay literature which is theology, academic (technical) and popular-modern. With these corpuses, we can study the effectiveness of the approaches and methods on three different setting.

The second component of this research that needs to be built and improved is the linguistic processing tool. This tool that will help the process of tokenizing or chunking, stemming and POS tagging will be based on two grammatical rules which is the 'pola' grammar algorithm done by [11] and CFG algorithm [10]. The reason behind this measure is to study and examined the effect of both Malay grammatical rules on the effectiveness rate of learning taxonomy method.

In studying the impact of different methods on Malay text, we proposed to adopt Cimiano's research methodology [3]. Cimiano's methodology is selected because most of the existing methodology and approach are not suitable for extracting Malay concept and relations between them. adopted a technique of reusing other sources (ontology) like WordNet, EuroWordNet or GermaNet to search for other words that might be synonymous and also to find relationship between concept which we have discussed the drawback of using WordNet in solving our problem.

Therefore, after developing the linguistic tools, we will continue by conducting an experiment using Formal Concept Analysis (FCA), a novel approach to the automatic acquisition of taxonomies or concept hierarchies from a Malay translated al-Qur'an which we already have in our repository. We will follow Harris' distributional hypothesis and model the context of a certain term as a vector representing syntactic dependencies which are automatically acquired from the text corpus with a linguistic parser developed in the first stage.

For evaluation and measurement, we will compare the learned concept hierarchies in terms of similarity with handcrafted reference taxonomies for two domains: Islamic Juriprudence (Fiqh) and thesis abstract from a computer science department. This is the common method in

benchmarking one research results in this field (please refere table 1). Other approaches and methods in taxonomy learning as shown in table 1 which have a higher degree of automaticity have to depend on a human expert to evaluate the learnt taxonomy.

We also will directly compare our approach with two other methods which is hierarchical agglomerative clustering as well as with Bi-Section-KMeans as an instance of a divisive clustering algorithm. Furthermore, we investigate the impact of using different grammatical rules and different measures weighting the contribution of each attribute as well as of applying a particular smoothing technique to cope with data sparseness in the Qur'anic data.

5.0 CONCLUSION

After having discussed the common approaches and methods in taxonomy learning and discussed a lot of work with respect to term extraction in English and also Malay, concept formation and concept hierarchy induction. However, we observe that there are no comparative work systematically analyzing different techniques and algorithms on Malay text. In advent of Semantic Web, ontology engineers from Nusantara (Malay-spoken region) need guidelines about the effectiveness, efficiency and trade-offs of different methods in order to decide which techniques to apply in which settings. We also have discussed above that each of the learning paradigms has advantages but also disadvantages. Thus, there is no paradigm that can produce optimal results. Therefore, we have proposed an experimental approach to study various approaches and methods to automatically acquire concept hierarchies from Malay texts. We summarize the main contributions of this project in the following: 1) Provide new corpus for ontology learning from Malay text; 2) A comparative study on the effect of using different Malay grammatical rule to approaches and methods; 3) This research addresses the lack of work in the field of concept hierarchy induction from Malay text by defining an evaluation methodology and systematically comparing different approaches with respect to the defined methodology.

BIBLIOGRAPHY

[1] P. Velardi, A. Cucchiarelli, and M. Pétit, "A Taxonomy Learning Method and its Application to Characterize a Scientific Web Community," *IEEE Transactions On Knowledge and Data Engineering*, vol. 19, pp. 180-191, 2007.

[2] A. Gomez-Perez, O. Corcho-Garcia, and M. Fernandez-Lopez, *Ontological Engineering*, 4th ed. New York: Springer-Verlag, 2005.

[3] P. Cimiano, A. Hotho, and S. Staab, "Learning Concept Hierarchies from Text Corpora Using Formal Concept Analysis," *Journal of Artificial Intelligence Research*, vol. 24, pp. 305–339, 2005.

[4] Wikipedia, "Taxonomy," Wikipedia Foundation inc., 2007.

[5] H. P. Alesso and C. F. Smith, *Developing Semantic Web Services*. Natick, Massachusetts: A K Peters, 2005.

[6] P. Cimiano, *Ontology Learning and population from Text*. Berlin Springer 2006.

[7] P. Cimiano, A. Hotho, and S. Staab., "Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text," presented at Proceedings of the European Conference on Artificial Intelligence (ECAI), 2004.

[8] A. Gómez-Pérez, D. Manzano-Macho, E. Alfonseca, R. Núñez, I. Blacoe, S. Staab, O. Corcho, Y. Ding, J. Paralic, and R. Troncy, "Deliverable 1.5: A survey of ontology learning methods and techniques," Next Web Generation, Leopold Franzens University of Innsbruck, Insitute of Computer Science, Innsbruck IST Project IST-2000-29243, 2003.

[9] A. Fatimah, "A Malay Language Document Retrieval System: An Experimental Approach and Analysis," in *Information Science*, vol. PhD. Bangi: Universiti Kebangsaan Malaysia, 1995, pp. 322.

[10] M. P. Hamzah, "Frasa dan Hubungan Semantik Dalam Perwakilan Pengetahuan:kesan Terhadap Keberkesanan Capaian Dokumen Melayu," in *Information Science*, vol. Phd. Bangi: Universiti Kebangsaan Malaysia, 2006, pp. 214.

[11] A. A. M. Juzaiddin, A. Fatimah, A. G. Abdul Azim, and M. Ramlan, "Pola Grammar Technique for Grammartical Relation Extraction in Malay Language," *Malaysian Journal of Computer Science,*, vol. 19, pp. 59-72, 2006.

[12] T. M. T. Sembok, "Word Stemming Algorithms and Retrieval Effectiveness in Malay and Arabic Documents Retrieval Systems," *Transactions on Engineering, Computing and Technology*, vol. 10, pp. 95-97, 2005.

[13] H. Schmid, "Probabilistic part-of-Speech Tagging Using Decision Tree," presented at International Conference on New Methods in Language Processing, Machester, UK, 1994.

[14] H. Schmid, "Lopar: Design and implementation. ," *Arbeitspapiere des Sonderforschungsbereiches*, vol. 340, 2000.

[15] M. S. Azhar, *Discourse-Syntax of "YANG" In Malay (Bahasa Malaysia)*. Kuala Lumpur: Dewan Bahasa dan Pustaka, 1988.

[16] K. Nik Safiah, "The Major Syntactic Structures of Bahasa Malaysia and their Implication of the Standardization

of the Language. , , ." vol. PhD. Ohio: Ohio University, 1975.

[17] C. K. Yeoh, "Interaction of Rules in Bahasa Malaysia," vol. PhD dissertation. Urbana-Champaign, : University of Illinois, 1979.

[18] H. O. Asmah, *Nahu Melayu Mutakhir*. Kuala Lumpur, : Dewan Bahasa dan Pustaka, 1980.

[19] K. Nik Safiah, F. M. Onn, H. H. Musa, and A. H. Mahmood, *Tatabahasa Dewan*. Kuala Lumpur: Dewan Bahasa dan Pustaka, 1993.

[20] A. Othman, "Pengakar perkataan melayu untuk sistem capaian dokumen," vol. MSc Thesis: National University of Malaysia, 1993.

[21] T. M. T. Sembok and P. Willett, "Experiments with n-gram string-similarity measure on malay texts," Universiti Kebangsaan Malaysia. . Bangi 1995.

[22] L. M. Iwanska, N. Mata, and K. Kruger., "Fully automatic acquisition of taxonomic knowledge from large corpora of texts," in *Natural Language Processing and Knowledge Processing*, L. M. I. a. S. C. Shapiro, Ed.: MIT/AAAI Press, 2000, pp. 335-345.

[23] W. Dolan, L. vanderwende, and S. Riichardson, "Automatically deriving structured knowledge bases from online dictionaries," presented at Proceedings of the Pasific Asociation for Computational Linguistics (PACLING), 1993.

[24] M. Hearst, "Automatic Acquisition of Hyponyms from large Text Corpora," presented at Proc. of 14th COLING, Nantes, France, 1992.

[25] M. Pasca, "Acquisiotion of categorised names entities for web search," presented at Proceedings of the Conference on Information and Knowledge Management (CIKM), 2004.

[26] O. Etzioni, M. Cafarella, D. Downey, S. Kok, and Popescu, "Web Scale Information Extraction in KnowitAll," presented at Proceedings of the 13th World Wide Web Conference, 2004.

[27] Z. Harris, *Mathematical Structure of Language*: Wiley, 1968.

[28] P. Velardi, R. Navigli, A. Cuchiarelli, and F. Neri, "Evaluation of OntoLean, a methodology for automatic population of domain ontologies.," 2005.

[29] D. Sanchez and A. Moeno, "Web Scale taxonomy learning," presented at Proceedings of the workshop on Extending and Learning Lexical Ontologies using Machine Learning Methods, 2005.

[30] M.-L. Reinberger, "Mining for Lexons:Applying unsupervised learning methods to create ontology bases.," presented at Proceedings of the International Conference on Ontologies, databases and Aplication of Semantics (ODBASE), 2003.

[31] W. Petersen, "A Set-Theoretical Approach for the Induction of Inheritance Hierarchies," in *Electronic Notes in Theoretical Computer Science* vol. 13 pages, 51 ed: Elsevier Science, 2001, pp. 13.

[32] S. Potter, *Language in the Modern World*. England: Penguin Books Ltd, 1968.

Table 1 Summary of Learning Taxonomy Methods From Text
Main Source: [8]

Researcher Name	Main Goals	Reuse Other	Main Technique	Source Use for	Evaluation
Velardi and Colleagues' Method 2007	To elicit a taxonomy	Yes	NLP Statistical approach	Domain text WordNet	Empirical measures And By experts
Cimiano's Approach 2006	To elicit a taxonomy	No	NLP Formal Concept Analysis	Domain Text	Empirical Measures & Gold Standard
Missikoff and colleagues' Method 2003	To build taxonomies and to fuse with an Existing ontology with	Yes	NLP Statistical approach ML techniques	Domain text WordNet	Expert
Park and Colleagues' Method	To learn new concepts	No	NLP Statistical approach Formal Concept Analysis	Domain Text	Empirical Measures
Alfonseca and Manandhar's method - 2002	To enrich an existing ontology with	Yes	Topic signatures Semantic distance	Domain text WordNet	Expert
Bachimont's method	To build a taxonomy	No	NLP techniques	Domain text	Expert
Xu and colleagues' Approach 2002	To learn concepts And relations between them	Yes	NLP techniques Statistical approach Text-mining techniques	Annotated text corpus WordNet	Expert
Khan and Luo's method 2002	To learn concepts	Yes	Clustering techniques Statistical approach	Domain text WordNet	Expert