International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 1, February 2018, pp. 1~10
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i1.pp1-10

Malay*IK*: An Ontological Approach to Knowledge Transformation in Malay Unstructured Documents

Fatimah Sidi[1], Iskandar Ishak[2], Marzanah A. Jabar[3]
[1,2]Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia
[3]Department of Software Engineering and Information System, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia

Article Info

Article history:

Received Feb 17, 2017
Revised Jul 13, 2017
Accepted Nov 29, 2017

Keyword:

Interrogative knowledge
Knowledge transformation
Ontology
Unstructured documents

ABSTRACT

The number of unstructured documents written in Malay language is enormously available on the web and intranets. However, unstructured documents cannot be queried in simple ways, hence the knowledge contained in such documents can neither be used by automatic systems nor could be understood easily and clearly by humans. This paper proposes a new approach to transform extracted knowledge in Malay unstructured document using ontology by identifying, organizing, and structuring the documents into an interrogative structured form. A Malay knowledge base, the Malay*IK* corpus is developed and used to test the Malay*IK*-Ontology against Ontos, an existing data extraction engine. The experimental results from MalayIK-Ontology have shown a significant improvement of knowledge extraction over Ontos implementation. This shows that clear knowledge organization and structuring concept is able to increase understanding, which leads to potential increase in sharable and reusable of concepts among the community.

Corresponding Author:

Fatimah Sidi,
Department of Computer Science,
Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia,
43400, UPM, Selangor, Malaysia.
Email: fatimah@upm.edu.my

1. INTRODUCTION

The difficulty of defining knowledge in unstructured documents is due to the paradox that knowledge resides in a person's mind and at the same time, it has to be captured, stored, and reported. For that, philosophers classify knowledge into knowing-that and knowing-how. Knowing-that is factual where data are stored in databases and facts can be recalled, processed, and disseminated. While knowing-how is actionable to do something, turning data into information and in turn into knowledge [1].

However, structured data represent only a little part of the overall organization of knowledge; in fact, the major part of this knowledge is incorporated in textual documents. For example, available business data are captured in text files that are not structured, e.g. memoranda and journal articles that are available electronically [2-4]. A large portion of the available information does not appear in structured databases but rather in collections of text articles drawn from various sources [5]. Thus, the main concern here is to dig knowledge from the available vast amount of textual documents.

The proposed ontological approach to knowledge transformation is based on interrogative structure [6] and conceptual modeling [7-11] approach. In transforming the extracted knowledge in unstructured document, "deep-level understanding" of complete sentences is extracted by identifying, organizing, and structuring the information into interrogative structured form. The "deep-level understanding" of complete

sentence refers to the understanding of a group of words in a complete sentence which, when they are written down, begin with a capital letter and end with a full stop, question mark, or exclamation mark.

The interrogative approach to knowledge extraction relies on data and conceptual modeling, as well as context and knowledge representation. Knowledge extraction supports the creation of: (1) knowledge; (2) relationship; (3) contextual information; and (4) representation of common languages. It gives aid in the transformation of extracted knowledge in an unstructured document into an interrogative structured form. The first issue to address corresponds to the need for a mechanism to identify knowledge from the sourced unstructured document in order to extract the knowledge. This is essentially in the interrogative knowledge identification, which identifies the type of document by separating text into knowledge, information or data and unifying it with personal components of values and beliefs. To identify knowledge, the approach of answering interrogatively is proposed to answer the question within the text in unstructured document.

The interrogative contextual information is derived from the incorporation of context and additional information annotation with context key facility. Context is an abstraction of the context factors, which are represented as concepts [12]. It is further exploited by [13] as contextual information, where information entered into the computer is tagged with context keys facilitating future retrieval using those keys. It may be any information that could be used to characterize the situation of an entity i.e. person, place, object [14]. For that, the interrogative contextual information is utilized to understand the process of making sense of information into knowledge and maintain the meaning of the information. This is to gain the interpretation of the identical knowledge by classifying the main point of the unstructured document interrogatively.

The rational to incorporate personal components towards the interrogative knowledge identification is as follows. According to [15], personal components have a powerful impact on organizational knowledge [16]. Assert that knowledge is a fluid mix of frame experience, values, contextual information, and expert insight. It originates in the mind of the knower to determine a large part of what the knower sees, absorbs, and concludes from his observations [17]. Stated that knowledge is a private and personal thing, which is intuitive and strongly linked to the user's values and beliefs. By manually transforming documents, values are embedded because humans read documents, extract the values of existing fields, and then enter the values into a user interface [18].

2. RESEARCH METHOD

This research proposes the Malay*IK*-Ontology model that is designed to transform extracted knowledge in Malay unstructured documents into an interrogative structured form based on interrogative knowledge identification as well as interrogative knowledge organization and structuring. The first step is to prepare the unstructured documents into an extension of plain text. The second step is to invoke the lexicon identifier that uses lexicon interrogative analysis matching rules of a specific corpus, which in this research, the Malay*IK*-Corpus. The lexicon identifier is used to identify and to extract knowledge in each of the complete sentences written in the Malay unstructured document. It is also used to extract interrogative lexical constructs from the individual unstructured document.

Next, the third step is to invoke the object recognizer that uses matching rules of object interrogative analysis in order to extract the ontological constructs from the interrogative lexical constructs. The object recognizer populates and maps the objects using ontology engineering, which is a mechanism of a knowledge structure to represent the concept and relationship of the abstract model on how people think about things in the world. Finally, the fourth step is to populate the database scheme by transforming the ontological constructs through connection between the ontology model and the object-relationship model.

The Malay*IK*-Ontology architecture (Figure 1) consists of three main components, which are *IKL*-Identifier, *IKO*-Recognizer, and *IKS*-OntologyDB. *IKL*-Identifier attempts to answer the question within the text interrogatively, and *IKS*-OntologyDB connects the ontology and object-relationship model to be populated into database.

The Protégé-Frames editor [19] is adopted in this research to structure and capture knowledge. It provides a full-fledged user interface and knowledge server to support users in constructing and storing frame-based domain ontologies, customizing data entry forms, and entering instance data. An object-based recognizer using interrogative knowledge approach or IKO-Recognizer [20], [21] is also adopted. In this research, the *IKO*-Recognizer maps the object interrogative analysis rule with ontology.

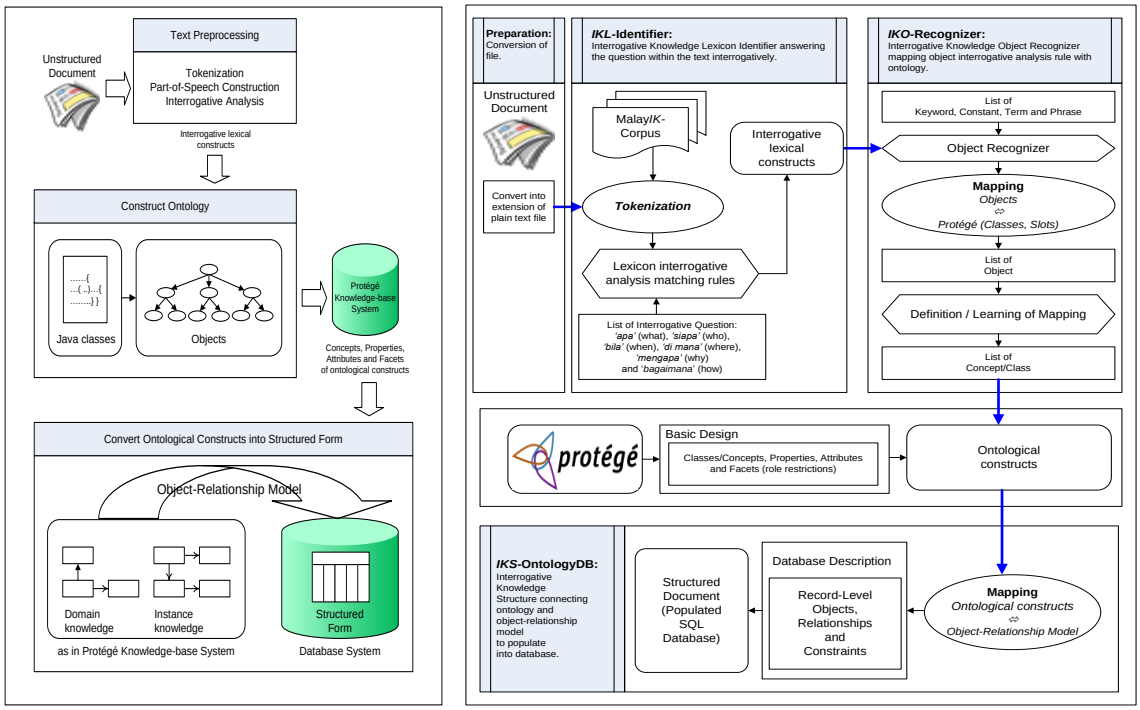

Figure 1. The MalayIK-Ontology Model(left) and System Architecture of MalayIK-Ontology(right)

2.1. *IKL*-Identifier

The Interrogative Knowledge Lexicon- (*IKL*-) Identifier is a lexicon identifier that uses lexicon interrogative analysis of *'apa'* (what), *'siapa'* (who), *'bila'* (when), *'di mana'* (where), *'mengapa'* (why), and *'bagaimana'* (how) in answering interrogative questions within the text in an unstructured document. The mechanism for the *IKL*-Identifier is to convert sentences into interrogative lexical constructs in the form of interrogative annotation.

Basically, the *IKL*-Identifier identifies the type of interrogative lexical constructs in each complete sentence within the Malay unstructured document by separating the text into knowledge, information or data. It is also responsible to tag the interrogative lexical constructs with interrogative contextual information, which is important to interpret the information into knowledge and maintain the meaning of the information in the Malay unstructured document. The processes of the *IKL*-Identifier are illustrated in Figure 2, which are tokenization, lexicon interrogative analysis, interrogative contextual information tagging, and phrases constructor tagging.

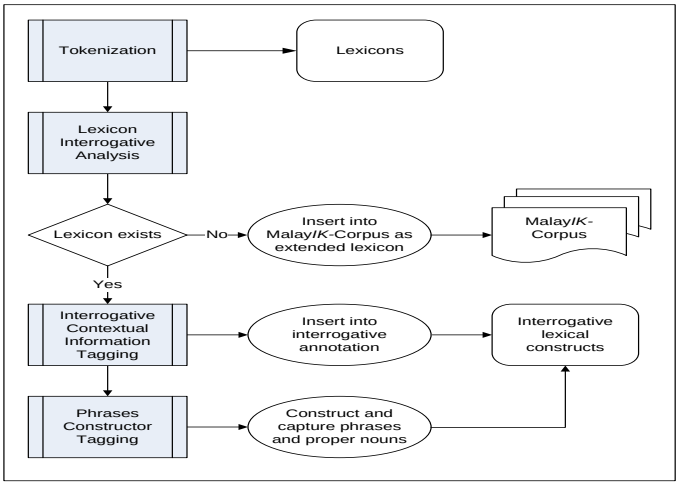

Figure 2. IKL-Identifier Processes

During tokenization, the text of unstructured document is segmented into sentences and tokenized into lexicons. Subsequently, the case format is defined, either the lexicon will hold digits, lower, upper, title or toggle cases. Each lexicon will then be assigned with automated serial number by lines, sentences, and token numbers. Next, during the Lexicon Interrogative Analysis, each lexicon is analyzed with lexicon interrogative analysis matching rules of the Malay*IK*-Corpus using the standard Data Manipulation Language (DML). DML is used to analyze, to check, and to insert the lexicon into interrogative annotation as interrogative lexical construct, should it exists. Any new lexicon that is analyzed will be inserted and defined in the Malay*IK*-Corpus.

Finally, the interrogative lexical constructs are used during Phrases Constructor Tagging. In this step, a phrase is constructed by putting together words based on interrogative annotation of the word. A phrase is a group of words, which contains an idea that forms a unit in which writing is part rather than a whole of a sentence. The words are divided depending on their use in a part of speech.

Malay sentence:
Brian Fielding Frost yang tercinta, umur 41, meninggal dunia pagi Selasa, September 30, 1998, disebabkan oleh kecederaan dialami dalam satu kemalangan kereta.

English sentence:
Our beloved Brian Fielding Frost, age 41, passed away Tuesday morning, September 30, 1998, due to injuries sustained in an automobile accident.

2.2. *IKO*-Recognizer

The *IKO*-Recognizer (Interrogative Knowledge Object Recognizer) performs matching and mapping object interrogative analysis rule of what/who/when/where/ why/how to extract ontological constructs [20], [21]. There are two major processes, object recognizer and mapping process. First, the object recognizer uses object interrogative analysis rules by utilizing Object-Oriented Programming (OOP) in order to conceptually organize the program around its data (objects/concepts). In this process, a number of object interrogative analysis rules and precondition language is pre-defined but users may manually define additional rules. Second, the following mapping process uses an ontology engineering approach, whereby objects that have been created by the object recognizer are accessible as plug-ins in the ontology system.

The main process in the IKO-Recognizer is the Object Interrogative Analysis Rules and the Precondition Language. Object interrogative analysis rules capitalize on Java OOP class encapsulation approach. For this, the object interrogative analysis rules use interrogative elements as the most upper class of the object. The structure and behaviors of the objects are implemented through (a) *Struktur Kata Nama Am* (Noun Structure) and (b) *Struktur Leksikon Semantik* (Semantic Lexicon Structure) in order to construct objects.

The first structure, which is the *Struktur Kata Nama Am* (Noun Structure), the object interrogative analysis rule is defined by combining the structure and behavior of an object with its inheritance and its conceptual modifiers of one or more subclasses in a hierarchical structure. The structure and behaviour of the object are defined by *'kata_masuk'* as tagged during the interrogative lexical construct earlier. The '*kata_masuk'* for *'penyelidik'* (investigator) is the grammatical information of 'kata nama am'. It is a noun of *'kata nama am orang'*, which refers to as a conceptual of *'Orang'* (People), and has the interrogative element of *'siapa'* (who). Hence, it inherits the general behaviour or properties of its parent *'siapa'* (who).

In the second structure, the *Struktur Leksikon Semantik* (Lexicon Semantic Structure), the object interrogative analysis rule uses the corresponding structure and behaviour of semantic lexicon that defines interrogative elements of *'bila'* (when), *'di mana'* (where), *'mengapa'* (why), and *'bagaimana'* (how). The semantic lexicons of *'bila'* (when) and *'di mana'* (where) correspond to the phrase or proper noun constructed after the semantic lexicon of the interrogative elements. The structure and behaviour of semantic lexicon *'bila'* (when) shows about the time at which an event take place. Whereas, the semantic lexicon of *'di mana'* (where) shows about the place something is in, or is coming from or going to.

However, the semantic lexicons of *'mengapa'* (why) and *'bagaimana'* (how) correspond to the predicate after the semantic lexicons of *'mengapa'* (why) and *'bagaimana'* (how). Reason being is to describe the meaning of the semantic lexicons and to give information about the sentence. The semantic lexicon of *'mengapa'* (why) talks about the reasons for something which introduces a relative. Whereas, the semantic lexicon *'bagaimana'* (how) explains the way in which something happens or is done and introduces a statement or fact. The objects of *'mengapa'* (why) and *'bagaimana'* (how) correspond accordingly to their definitions of interrogative element.

2.3. *IKS*-OntologyDB

The *IKS*-OntologyDB is a process of exporting the ontology structure into a database. The metadata of the information regarding the relationships, properties, attributes, and facets of the class structure are created in the ontology system and are exported into Microsoft Access. The exportation is done by using the facility provided by Protégé by selecting the option of Export to HTML format. The transformation of the knowledge-based system created via the Protégé to the database management system by using HTML format. The HTML information is used to create attributes and constraints in the Microsoft Access. The table is created according to the definition and declarations of the SQL schema and ontology declaration of the Protégé knowledge-based system. This is shown in Table 1.

Table 1. Ontology versus SQL Declaration Properties

Ontology Declaration	SQL Declaration
Class	Table
Slot	Field Name
Property	Data Type
Facet	Constraint
Cardinality	Cardinality

The components of ontology and conceptual model are basically equivalent in terms of concepts and entities, relationships, and attributes. The ontological constructs generated are mapped with the Object-oriented System Model (OSM) established by [7-11]. It is used by the object-relationship model to describe the data interest which includes relationships, lexical appearances and context keywords. Besides, it is used to structure the data identified and extracted and populate them into database scheme.

In general, the relevant knowledge about an object set is represented by a colon (:) after an object-set name which denotes that the object set is a specialization. For example, the lexical object set of Death Date: Date, where date describes the string patterns of interrogative element of *'bila'* (when). For the lexical object set of Deceased Name: Name and Relative Name: Name, name is matched by recognizing the string patterns of proper nouns interrogative element of *'siapa'* (who). The context keywords indicate the presence of an object in an object set. For example, *'kematian'* (died) and *'meninggal dunia'* (passed away) are the context keywords for Death Date; *'pengebumian'* (buried) is a context keyword for Interment.

2.4. Malay*IK* Corpus

This research uses the Malay*IK*-Ontology based on interrogative approach. While most approaches of text processing as discussed in [22] use NLP or information extraction to select the set of keywords or phrases to be analyzed, ontology approach is able to avoid mislead in the "vocabulary problem" which leads to spurious results. By establishing a fixed set of general concepts ("People", "Location", "Things") with the entry of word answering the question interrogatively ("People", "Location", "Things" refer to 'who', 'where', 'what' respectively), the vocabulary used in the rule mapping phase may be controlled.

The most important attribute is the grammatical information of lexicon entry to answer the question of the lexicon grammatical information interrogatively besides the root word (lexicon). The Malay*IK*-Corpus is a Malay language corpus where the Malay dictionary of *Kamus Dewan* [23, 24] and the dictionary of root words act as important secondary controls of the lexicon entries. It also refers to the dictionary of *Kamus Imbuhan Bahasa Melayu* [25], *Kamus Dwibahasa Oxford Fajar* [26], and *Kamus Komprehensif Bahasa Melayu* [27]. The lexicons entries are manually inserted in the database using standard DML of the related database.

In order to create a general purpose corpus for Malay language, the Ahmad's and Abdullah's stop words [23], [25] are included which indicate pronoun, auxiliary verb, adverb, predicate, preposition, negative, conjunction, relative and determinant.

Table 2 presents examples of words entry extracted from Malay*IK*-Corpus in a table format (by columns and rows). The header row of Table 2 refers the attributes of corpus by columns. The rest of the rows are examples of words entries for *'rumah'* (house), *'sejak'* (since), *'penyelidik'* (researcher), *'di'* (at), *'kerana'* (because) and *'dengan'* (with). It answers the question of interrogative of *'apa'* (what), *'bila'* (when), *'siapa'* (who), *'di mana'* (where), *'mengapa'* (why), and *'bagaimana'* (how) respectively.

Basically, the grammatical information of *'rumah'* and *'penyelidik'* is noun *('kata nama am')* but are classified as different category. The word *'rumah'* (house) falls under categorization of 'Things' which answers the interrogative question of 'what'. While *'penyelidik'* (researcher) falls under categorization of 'People' which answers the interrogative question of 'who'.

Table 2. Examples of MalayIK-Corpus

kata dasar	Perkataan	kata masuk	elemen interogatif	Status
rumah	rumah	kata nama am benda	apa	1
sejak	sejak	kata sendi nama masa	bila	2
selidik	penyelidik	kata nama am orang	siapa	1
di	di	kata sendi nama tempat dan arah	di mana	2
kerana	kerana	kata hubung pancangan	mengapa	2
dengan	dengan	kata sendi nama bersama-sama	bagaimana	2

3. RESULTS AND ANALYSIS

The first question that arises in designing the Malay*IK*-Ontology is to check whether the constant/keyword recognizer to extract and structure data of Ontos can be applied to Malay unstructured documents. The next question is to check whether the Malay*IK*-Ontology can identify knowledge as well as data to be extracted and structured are equivalent or better than Ontos. Furthermore, the knowledge or data obtained needs to be checked for its validity. This is to prove that the Malay*IK*-Ontology works effectively in extracting and structuring data as compared with Ontos. Therefore, following are the steps taken to perform the experiment.

The accuracy of Ontos is measured by the numbers of data extracted between the English and Malay obituaries. The accuracy of Malay*IK*-Ontology is measured by numbers of knowledge or data extracted. When applying Ontos on English and Malay obituaries, three tables are created based on the obituaries ontology, which are DeceasedPerson, Viewing, and DeceasedPersonRelationshipRelativeName. For Malay*IK*-Ontology, the table DeceasedPersonRelationshipRelativeName is used to compare the translation of Malay language for Relationship. An example of data extracted for both Ontos and Malay*IK*-Ontology are listed in Table 3 for DeceasedPerson.

Table 3. List of Data Extracted for Table DeceasedPerson

	Extraction	Manual Extraction	Ontos		Malay*IK*-Ontology
	Language of Obituary	English	English	Malay	Malay
Facts/Attributes	DeceasedPerson	1002	1002	1002	1002
	DeceasedName	Lemar K. Adamson	Lemar K. Adamson	Lemar K. Adamson	Lemar K. Adamson
	Age	84	84	199	84
	DeathDate	9/30/1998	9/30/1998	9/30/1998	9/30/1998
	BirthDate	6/12/1914	6/12/1914	6/12/1914	6/12/1914
	Funeral	5002	5002	5002	5002
	FuneralDate	10/5/1998	10/5/1998	-	10/5/1998
	FuneralAddress	Silverbell Ward, 1540 E. Linden	1540 E. Linden	1540 E. Linden	Silverbell Ward, 1540 E. Linden
	FuneralTime	10:00 AM	10:00	10:00	10:00 AM
	Interment	7002	7002	7002	7002
	IntermentAddress	City Cemetery	236 S. Scott	236 S. Scott	City Cemetery
	IntermentDate	-	-	-	-

(-) No data extracted

3.1. Analysis of Ontos on English and Malay Obituaries

The results of Ontos being applied on English and Malay obituaries are shown in Table 4 and Table 5. This table shows the counted number of facts (attributes values) in the test-set documents of English and Malay obituaries. They [7-11] are consistent with their implementation, which only extracts explicit constants. A string is counted as correct, if the constant extracted occurs in the text. With this understanding, counting is basically straightforward. Due to their name lexicon is incomplete and because of their name-extraction expressions are not rich; sometimes parts of a name are missed or a single name were split into two. For these cases, they list the count after + in the Declared Correctly column.

Partial names also caused most of the problems for the large number of incorrectly identified relatives. With a more accurate and complete lexicon coupled with richer name-extraction expression, they believe they can achieve much higher precision.

As anticipated, the experiment to check the constant/keyword recognizer of Ontos on Malay obituaries does not produce the same results as English obituaries for numbers of facts generated. However, the results show that the DeceasedPerson, DeceasedName, BirthDate and DeathDate generate 100% recall and precision. Facts generated are classified as nonlexical and lexical objects set. The nonlexical and lexical objects set are described in what they defined as data frames. A data frame describes the string patterns for its constants.

For that, results of lexical objects sets such as counted number of facts for the IntermentDate, IntermentAddress, ViewingDate, ViewingAddress, Relationship and RelativeName generate 0% recall and precision listed in Table 4. This is due to the context keywords in Malay obituaries being translated. However, non-lexical object sets such as DeceasedPerson are always generated for an obituary record and consequently the results for that sets are 100% recall and precision. They represent non-lexical object sets by surrogate identifiers which are generally easier to identify correctly. This shows Ontos can be applied on Malay obituaries for data frame of non-lexical object sets by surrogate identifiers.

Table 4. Results of Ontos on English and Malay Obituaries

Facts	Number of Facts in Source (N)	Number of Facts Declared Correctly + Partially Correct (C)		Number of Facts Declared Incorrectly (I)		Recall Ratio (%) (C/N)		Precision Ratio (%) (C/C+I)	
		E	M	E	M	E	M	E	M
DeceasedPerson	3	3	3	0	0	100	100	100	100
DeceasedName	3	3	3	0	0	100	100	100	100
Age	3	3	3	0	3	100	0	100	50
BirthDate	3	3	3	0	0	100	100	100	100
DeathDate	3	3	3	0	0	100	100	100	100
			Funeral						
FuneralDate	3	3	0	0	0	100	0	100	0
FuneralAddress	3	0	0	3	2	0	0	0	0
FuneralTime	3	2	1	1	1	67	33	67	50
			Interment						
IntermentDate	0	0	0	2	0	0	0	0	0
IntermentAddress	3	0	0	3	3	0	0	0	0
			Viewing						
ViewingDate	1	0	0	0	0	0	0	0	0
ViewingAddress	3	0	0	0	1	0	0	0	0
BeginingTime	3	2	1	0	1	67	33	100	50
EndingTime	3	3	1	0	0	100	33	100	100
			RelationshipRelativeName						
Relationship	8	8	0	0	0	100	0	100	0
RelativeName	28	6	0	9	0	22	0	40	0

(E) English, (M) Malay

3.2. Analysis of Malay*IK*-Ontology on Malay Obituaries

The results of Malay*IK*-Ontology being applied with Malay obituaries to determine its ability to identify and extract data as compared with Ontos are listed in Table 5. Table 5 shows the results obtained are compared with the results of the first experiment on Ontos. The purpose of comparison is to check the validity of data identified and generated by the Malay*IK*-Ontology. In this second experiment, results of both nonlexical (DeceasedPerson) and lexical objects (Funeral, Interment, Viewing and RelationshipRelativeName) of Malay*IK*-Ontology generate 100% recall and precision except for ViewingAddress which generates 67% recall and precision. This is due to the location keyword indicating

address for the lexical object of ViewingAddress in Malay obituaries being translated as something done by a person. Whereas, the lexical object for IntermentDate definitely generates 0%, this is because there is no interment date specified in the obituaries. Hence, this shows that there is an improvement in extracting and structuring data of Ontos lexical objects by the Malay*IK*-Ontology.

Table 5. Results of Ontos and MalayIK-Ontology

Facts	*N*	Number of Facts Declared Correctly + Partially Correct *(C)*			Number of Facts Declared Incorrectly *(I)*			Recall Ratio (%) *(C/N)*			Precision Ratio (%) *(C/C+I)*		
		Ontos		MO	Ontos		MO	Ontos		MO	Ontos		MO
		E	M	M	E	M	M	E	M	M	E	M	M
DeceasedPerson	3	3	3	3	0	0	0	100	100	100	100	100	100
DeceasedName	3	3	3	3	0	0	0	100	100	100	100	100	100
Age	3	3	3	3	0	3	0	100	0	100	100	50	100
BirthDate	3	3	3	3	0	0	0	100	100	100	100	100	100
DeathDate	3	3	3	3	0	0	0	100	100	100	100	100	100
Funeral													
FuneralDate	3	3	0	3	0	0	0	100	0	100	100	0	100
FuneralAddress	3	0	0	3	3	2	0	0	0	100	0	0	100
FuneralTime	3	2	1	3	1	1	0	67	33	100	67	50	100
Interment													
IntermentDate	0	0	0	0	2	0	0	0	0	0	0	0	0
IntermentAddress	3	0	0	3	3	3	0	0	0	100	0	0	100
Viewing													
ViewingDate	1	0	0	1	0	0	0	0	0	100	0	0	100
ViewingAddress	3	0	0	2	0	1	1	0	0	67	0	0	67
BeginingTime	3	2	1	3	0	1	0	67	33	100	100	50	100
EndingTime	3	3	1	3	0	0	0	100	33	100	100	100	100
RelationshipRelativeName													
Relationship	8	8	0	8	0	0	0	100	0	100	100	0	100
RelativeName	28	6	0	28	9	0	0	22	0	100	40	0	100

(*N*) Number of Facts in Source, (E) English, (M) Malay, (MO) Malay*IK*-Ontology

4. CONCLUSION

The main objective of this research is to propose a new approach to transform extracted knowledge in Malay unstructured document by identifying, organizing, and structuring them into interrogative structured form. In order to achieve this objective, an approach is established through the Malay*IK*-Ontology approach. Based on the results, the annotation of interrogative contextual information tagged in interrogative lexical constructs improves the data extraction. The annotation of interrogative contextual information is annotated with interrogative and grammatical information of the lexical constructs. For example, the lexical object set of BirthDate, DeathDate, FuneralDate, and ViewingDate generate precision of 100% which also generate 100% precision on Ontos.

This improvement is due to the implementation of annotation interrogative contextual information which is tagged with interrogative element of *'bila'* (when) which describes about the string patterns of time at which things happened. For lexical object set of DeceasedName and RelativeName which also generate precision of 100%, as the name is matched by recognizing the string patterns of proper nouns which is tagged with interrogative element of *'siapa'* (who). Besides, phrases or proper nouns based on interrogative annotation of the lexicon are also annotated in the lexical constructs.

5. ACKNOWLEDGEMENT

The work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS 03-12-10-999FR).

REFERENCES

[1] Spiegler, I.: Technology and Knowledge: Bridging a "Generating" Gap, *Information & Management*, vol. 40(6), pp. 533-539, 2003.
[2] T. K. Das and P. M. Kumar, "Big Data Analytics: A Framework for Unstructured Data Analysis", *International Journal of Engineering and Technology*, Vol. 5, pp. 153-156, 2013
[3] K. Cukier, "The Economist, Data, data everywhere: A special report on managing information", 2010, February 25, retrieved from http://www.economist. com/node/15557443
[4] N. Khan, *et al.*, "Big Data: Survey, Technologies, Opportunities, and Challenges", *The Scientific World Journal*, vol. 2014, pp. 1-18, 2014.
[5] R. Feldman, *"Mining Unstructured Data"*, in International Conference on Knowledge Discovery and Data Mining 5th ACM SIGKDD, August, 1999.
[6] E. J. Quigley and A. Debons, "Interrogative theory of information and knowledge", in *ACM SIGCPR conference on Computer personnel research*, pp. 4-10, 1999, pp. 4-10.
[7] D. W. Embley, et al., *"A Conceptual-Modeling Approach to Extracting Data from the Web"*, in 17th International Conference on Conceptual Modeling, 1998, pp. 78-91.
[8] D. W. Embley, *"Ontology-based Extraction and Structuring of Information from Data-rich Unstructured Documents"*, in 7th International Conference on Information and Knowledge Management, 1998, pp. 52-59.
[9] D. W. Embley, *et al.*, "Conceptual-Model-based Data Extraction from Multiple-Record Web Pages, Data and Knowledge Engineering", *Data & Knowledge Engineering Journal*, vol. 31(3), pp. 227-251, 1999.
[10] D. W. Embley, et al., *"Record-Boundary Discovery in Web Documents"*, in Proceeding of the 1999 ACM SIGMOD International Conference on Management of Data, 1999, pp. 467-478.
[11] D. W. Embley, *"Toward Semantic Understanding: An Approach Based on Information Extraction Ontologies"* in 15th Australasian Database Conference, 2004, pp. 3-12.
[12] B. N. Schilit and M. M. Theimer, "Disseminating Active Map Information to Mobile Hosts", *IEEE Network*, vol. 8(5), pp. 22-32, 1994.
[13] Lamming, M.G. and Newman, W.M., *"Activity-based Information Retrieval: Technology in Support of Personal Memory"*, in 12th World Computer Congress on Personal Computers and Intelligent Systems - Information Processing, 1992, pp. 68-81.
[14] G. D. Abowd, et al., *"Towards a Better Understanding of Context and Context-Awareness"*, in 1st International Symposium on Handheld and Ubiquitous Computing 1999, pp. 304-307.
[15] I. Nonaka, "A Dynamic Theory of Organizational Knowledge Creation". *Organization Science*, vol. 5(1), pp. 14-37, 1994.
[16] T. H. Davenport, *et al.*, "Working Knowledge: How Organizations Manage What They Know", *Harvard Business School Press*, 2000.
[17] F. J. Varela, *et al.*, "The Embodied Mind", *MIT Press*, 1992.
[18] R. Virk, "Transforming Unstructured Content into "Meaningful" XML", 2004, http://www.dmreview.com/whitepaper/WID413.pdf.
[19] N. F. Noy, and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology", *Knowledge Systems Laboratory*, Technical Report KSL-01-05, 2001.
[20] I. Ishak, *et al.*, "Object Recognizer For Organizing And Structuring Unstructured Documents Using Interrogative Knowledge", *Journal of Theoretical & Applied Information Technology*, vol. 83, pp. 442-450, 2016.
[21] F. Sidi, F. and M. A. Jabar, "Interrogative Knowledge Organization and Structuring from Unstructured Documents". *Knowledge Management and Innovation: A Business Competitive Edge Perspective*, pp. 1028-1032, 2010.
[22] M. Shamsfard and A. A. Barforoush, "The State of the Art in Ontology Learning: A Framework for Comparison", *Knowledge Engineering Review*, vol. 18(4), pp. 293-316, 2003.
[23] Dewan Bahasa Perpustakaan, "Kamus Dewan Edisi Ketiga", *Dewan Bahasa dan Pustaka*, 2002.
[24] Dewan Bahasa Perpustakaan, "Kamus Dewan Edisi Keempat", *Dewan Bahasa dan Pustaka*, 2005.
[25] H. M. Ali, "Kamus Imbuhan Bahasa Melayu Edisi Kedua", *Penerbit Fajar Bakti Sdn. Bhd.*, 1993.
[26] J. M. Hawkins, "Kamus Dwibahasa Oxford Fajar Edisi Ketiga", *Penerbit Fajar Bakti Sdn. Bhd.*, 2001.
[27] A. Othman, "Kamus Komprehensif Bahasa Melayu", *Penerbit Fajar Bakti Sdn. Bhd., a subsidiary of Oxford University Press*, 2005.

BIOGRAPHIES OF AUTHORS

Fatimah Sidi holds a PhD in Management Information Systems from Universiti Putra Malaysia. She received her Bachelor and Masters degree in Computer Science from the same university. She has twenty years experience as System Analyst in System Development and Enterprise Database Management. Her research interest spans around Knowledge and Information Management Systems, Data and Knowledge Engineering, as well as Database and Data Warehousing

Iskandar Ishak received his BIT (Hons) from Universiti Tenaga Nasional in 2002. He received Master of Technology in Information Technology in the year 2003 from the Royal Melbourne Institute of Technology University and completed his PhD in Computer Science in 2011 at the Universiti Teknologi Malaysia. He is now a Senior Lecturer at the Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia since 2011. His research interest lies in the area of Database System, Information Systems, Big Data, Peer-to-Peer and Mobile Computing and Information Retrieval

Marzanah A. Jabar graduated from Universiti Kebangsaan Malaysia with BSc. in Quantitative Studies in 1983. She received her MSc. in Computer Science in the year 2001 and her PhD in Management Information System in the year 2007, both from Universiti Putra Malaysia. She has over twenty years experience in leading information systems development. Her research interest lies in the area of Software and Knowledge Engineering, Knowledge Management Systems, and strategic use of Information Systems and Technologies