Developing an Automatic Ontology Constructor for Bahasa Indonesia Using Cognitive Approach: A Proposal

Gloria Virginia[1], **Hung Son Nguyen**[2]

virginia@icm.edu.pl, son@mimuw.edu.pl

ABSTRACT

There are great challenges in computer linguistic and information retrieval fields which process Bahasa Indonesia. Taking advantage from implementation of Linear Model life cycle, an automatic ontology constructor (OC) is going to be generated. The natural language in emails together with the exhaustive cognitive approach argues to be essential in the process of OC generation that one may achieve deep linguistic analysis. An ontology-based information retrieval system for Indonesian choral community is going to be developed using 3,000 emails of Indonesian Choral Lovers (ICL) mailing list as a tool of evaluation. Performance measure of information retrieval (recall and precision) and qualitative measure of ontology (consistency, completeness, and conciseness) are going to be used to find out the OC effectiveness through examining the automatic-thesaurus effectiveness. The associative thesaurus generated manually and automatically will also enrich the IndonesianWordNet being struggled.

Keywords : *Ontology, thesaurus, cognitive approach, text retrieval system*

INTISARI

Memproses Bahasa Indonesia merupakan tantangan besar dalam area *computer linguistic* dan *information retrieval*. Pada studi ini, daur hidup Linear Model akan diimplementasikan dalam rangka menghasilkan sebuah konstruktor ontologi otomatis (OC). Bahasa alami, penggunaan email, dan pendekatan kognitif dianggap sebagai hal penting dalam proses analisa linguistik. Sebuah sistem temu kembali berbasis ontologi khusus bagi komunitas paduan suara Indonesia akan dibangun sebagai alat evaluasi, dimana 3,000 email yang diambil dari milis *Indonesian Choral Lovers* (ICL) akan menjadi data utama. Efektifitas OC akan diukur menggunakan dua alat ukur, yaitu alat ukur performa dari sistem temu kembali (*recall* dan *precision)* dan alat ukur kualitatif dari ontologi *(consistency, completeness,* and *conciseness)*.

Kata Kunci : Ontologi, tesaurus, pendekatan kognitif, sistem temu kembali

INTRODUCTION

Indonesian Choral Lovers (ICL) is a Yahoo! Groups for Indonesian whom loves to sing in a choir. There are more than a thousand people subscribed, including those who considered as choral experts in Indonesia. Since it was made in March 2000, the number of posted email increased significantly, mainly in the second year (from 4 emails to 1,350 emails) and these last three years (from 840 emails in 2005 to 1,019 emails in 2006, 1,525 emails in 2007, and 1,886 emails in 2008). Recently, it is more than 11,000 emails and the content are varied from merely information to question and discussion.

[1] Faculty of Mathematics, Informatics, and Mechanic, University of Warsaw, Poland

[2] Faculty of Mathematics, Informatics, and Mechanic, University of Warsaw, Poland

Since the author's membership into this group in 2004, numerous information scattered and buried under the emails is observed. One discernible indicator is many topics have been questioned and discussed continuously while a new member signed up. The repetition retards the learning process in the community and makes the discussion boring frequently. The possibility to receive qualified answer also decreases, especially when it is related to the common topic of discussion, because people seems to get tired to explain it time after time. It becomes more interesting while many members prefer posting a question in the form of natural language and rely on response from others to using the available search engine.

A type of question-answering machine for choral lovers seems to be a great idea to encounter such problem where an answer-mail is generated automatically after a question-mail is posted by a user. However, an accurate answer of a query could be arranged while the system has relevant documents retrieved by a retrieval system. It could be realized while appropriate knowledge is attached into the system.

It seems that an ontology-based text retrieval system should be developed at the first stage of the intended question-answering machine. Regarding that there will be another process afterwards which is arrangement of an answer, system effectiveness and efficiency should take into consideration. Moreover, the continual increasing number of emails in mailing list leads to expectation of an automatic ontology constructor in the system. This ontology constructor should have ability to deal with natural language and semantic analysis.

For this, particular study for Bahasa Indonesia using cognitive approach and emails as the corpus argues to be important in order to achieve deep linguistic analysis. Available methods of other language (e.g English) cannot be simply adapted regarding the specific morphological rules of Bahasa Indonesia. For example, affixes of Bahasa Indonesia consists of [1]:

- prefix (e.g. "penyanyi" [a singer] has the root "nyanyi" [to sing] and prefix "pe"),
- suffix (e.g. "nyanyian" [a song] comes from the root "nyanyi" [to sing] and suffix "an"),
- infix (e.g. "gemetar" [shaky] is derived from the root "getar" [a trill] by inserting the infix "me" between the "g-' and "-etar" of "getar"),
- confix (e.g. "pelatihan" [exercising] has the root "latih" [to exercise], prefix "pe", and suffix "an").

Section 2 describes an overview of relevant issues to information retrieval with embedded knowledge. While research in this area has been growing in big strides, section 3 will show that the Indonesian research community is working to move progressively. A proposed system and the methodology being used are discussed in section 4 and 5 consecutively before summary and suggestions of potential studies given in section 6. Detail explanation on implementation of Linear Model is available at section 5.1, while at section 5.2 and 5.3 will describe the ontology constructor and its evaluation.

ONTOLOGY-BASED TEXT RETRIEVAL SYSTEM

An ontology is an explicit specification of a conceptualization [2]. Two important features of a model to be considered as an ontology are it is sharable and consensual knowledge model agreed by a community [3]. Thesauri were classified as a lightweight ontology in a continuous line of ontology spectrum by [4] which provide some additional semantics between terms. It gives information such as synonym relationship but do not provide an explicit hierarchy.

Indication that the ontology-based system can be used to enhance system performance showed in many studies. Frequently Asked Questions (FAQ) Finder system of [5] use natural language question interface to access distributed FAQ files and returns the given answers. They showed that a combination of techniques from information retrieval and natural language processing work better than any single approach. A study in [6] generated ontology and implemented it in semantic search engine in institutional repository of scholarly literature in Malaysia to find conceptual relations between documents and retrieve related articles in a more efficient way. The integration process seems to work with small data that they were experimenting with. A study in WebDoc (a system that classifies Web documents according to the Library of Congress classification system) was promising although they worked on a relatively small training set and without the semantic tags on noun phrase and other contextual information [7]. They used thesaurus and certainty factor (CF) for index generation. The CF was derived automatically by combining multiple sources of evidence (good hits and bad hits) in order to take into consideration the quantitative aspect of the relationships between terms.

An ontology is usually hand crafted by domain experts but it is realized to have major problems from construction to maintenance, mainly as high initial cost, its tendency to evolve rapidly over time [8] and change between different applications [9]. In recent years, there has been an increasing study about learning or adapting ontology dynamically. A study in [10] used Protégé plug in to bootstrap a domain-specific ontology from a relevant text corpus of neurology while [11] used conceptual clustering algorithm, COBWEB, to automatically generate class hierarchies as a basic ontology represented using Resource Description Framework (RDF) Schema in a Smart Radio online song recommendation application.

CHALLENGES

Indonesia is an archipelago country that has 17,508 islands[3] and population over 240 million[4]. It is divided into 33 provinces and recognize specific ethnics groups comes from certain province which has particular ethnic language as their mother tongue. However, Bahasa

[3] http://www.indonesia.go.id/. Accessed on 20 September 2010.

[4] https://www.cia.gov/library/publications/the-world-factbook/geos/id.html. July 2009 estimation. Accessed on 20 September 2010.

Indonesia becomes the official language of Indonesia and most Indonesians are proficient in using the language for communication because it is taught at all school.

The Asosiasi Penyelenggara Jasa Internet Indonesia (Indonesian Internet Service Provider Association) recorded that Internet subscribers and users of Indonesia has been growing rapidly[5]. Even The World Factbook of CIA put Indonesia at the eleventh place based on the number of Internet users with 30 million users[6]. These facts should encourage researches such as computer linguistic and information retrieval for Bahasa Indonesia which in fact has not been extensively investigated. Limited number of scholarly article in these areas is an indicator.

Considerable effort is showed by Indonesian research community since mid of 1990s [12]. Latest study of Indonesian morphological analysis is a morphological analyzer for Bahasa Indonesia developed by [13] which is an extension version of former works (i.e. [14] and [15]). The implementation of morphotactic and morphophonemic rules in the study provided richer semantic, lexical and grammatical information of words. A study in [16] evaluated a number of stemmer and concluded that confix-stripping stemmer (cs stemmer) which is an improvement of [17] is a highly effective tool for Bahasa Indonesia.

An ongoing study is an IndonesianWordNet (IWN) development [18]. Instead of using merge-model, it uses expand-model to map the common base concepts between PrincetonWordNet (PWN) [19] synset and Kamus Besar Bahasa Indonesia (KBBI)[7] sense definition by involving large numbers of human annotators through web-based application. The initial experiments showed that the mappings can be used to construct a fairly reliable Indonesian WordNet.

IWN development is argued to be significant in computer linguistic and information retrieval field of study because it will support numbers of researches. However, the study being conducted is suffer and danger in semantic particularly because of an inherent asymmetry mapping process between PWN and IWN. The quality of the contents and structure of IWN yielded should be considered carefully, although the study showed the degree of agreement between several annotators are fairly tolerated statistically. A large number of human annotators were not sufficient in order to have high reliability, because we never know their capabilities in language and their background of intention while doing the evaluation. On the study, the number of annotators being used is not comparable with the number of words evaluated which is thousands hence unevaluated words are highly probable.

A few numbers of relevant studies about text retrieval applications in Bahasa Indonesia are identified but none in ontology constructor exclusively. A study in [20] tried to apply the use

[5] http://www.apjii.or.id/dokumentasi/statistik.php. Updated on December 2007. Accessed on 20 September 2010.

[6] https://www.cia.gov/library/publications/the-world-factbook/geos/id.html. Date of information: 2008. Accessed on 20 September 2010.

[7] KBBI is a comprehensive dictionary of Bahasa Indonesia developed by the Indonesian government's centre for language development (Pusat Bahasa) of Ministry of Education.

of Latent Semantic Indexing and Semi-Discrete Matrix Decomposition in Bahasa Indonesia information retrieval system using student research documents at Computer Science Department of Institut Pertanian Bogor (IPB). Implementation of inference network in Indonesian news articles was tried by [21]. In [22], a study of spoken query-based Indonesian information retrieval was presented.

PROPOSED MODEL

An ontology-based text retrieval system for Indonesian choral should be the first system acquired in order to arrive at the more complex intended system which is question-answering machine for choral lovers. Figure 1 is the use-case diagram of the proposed model for this study.

The study could be seen as a labor to improve and enrich the IWN work. Instead of using common base concept, it focuses in choral domain and crucial words in ICL emails. The expectation is the IWN will improve broader and deeper on specific subject. In spite of acknowledgement that good classification and concept coverage do not guarantee effective retrieval [23], experts in choral are preferred to be involved particularly in annotation task in order to pursue the quality of annotation. With limited number of experts, discussion of topic distinction could be facilitated in proper manner and reduce complexity of debate between experts.

Cognitive science is the study of mental representations and computations and of the physical systems that support those processes [24]. This approach is performed in revealing the extraction process of important words and semantic relationship between them.

The manually process of choral thesaurus creation should benefit generation of automatic ontology constructor. Taking advantage of the growing corpus, the constructor should revive the choral thesaurus particularly. Generation of heuristic rules from the experts is highly plausible and will be fed as the system knowledge that cooperates with the constructor work. Finally, indexing process and query refinement should be boosted by the existence of continues updated-ontology. This should lead to better relevant emails retrieved.

Availability of IWN should give much advantage on the study to investigate the achievement of developed constructor. However, Indonesian lexicon[8] will be appropriate on this initial study.

In accordance with Text REtrieval Conference (TREC) format, test collections consist of three parts which are a set of documents, a set of information needs, and relevance judgments [25]. The test collection needed in the study (e.g. email corpus) is not available, hence it should be developed from ICL mailing list.

[8] This Indonesian lexicon is used by IR Lab of University of Indonesia in their studies these years.

In this study, the front-end of information retrieval system being developed is a search engine, while at back-end there is choral thesaurus as system knowledge that will be used in indexing and query refinement. Users are able to input their queries through natural language and the output will be a list of relevant emails in the corpus. However, the focal point of this study is the development of automatic ontology constructor while the search engine developed is going to be a tool in evaluating the constructor performance. Details explanation about ontology constructor is available on section 5.2.

LINEAR MODEL

A text retrieval system based on ontology could be seen as a combination between expert system and retrieval system. While the cognitive approach argues to be compulsory, the Linear Model that has been successfully used in a number of expert system projects [26] seems to be fit implemented. This system life cycle consists of six stages which are planning, knowledge definition, knowledge (and system) design, code and checkout, knowledge verification, and system evaluation. Table 1 presents the summary of Linear Model implemented during the study, including the important tasks of each stage and its main output.

The important task on the first stage is corpus preparation, which are emails retrieved from ICL mailing list since March 2000. Instead of using entire emails, the 3,000 first emails are considered to be appropriate as the ICL-corpus for this initial study. We cannot choose 3,000 emails randomly by remembering that email is a kind of communication tool, hence there are continuous conversations between writers. While the raw data is ready, choral experts commit annotation task which is annotating topic(s) on each email as well as listing the main words describing its topic(s). The result of annotation process are annotated corpus as well as the TREC-like test collection where the documents will be used as a corpus of the retrieval system; the information needs and relevance judgments will be used in system evaluation.

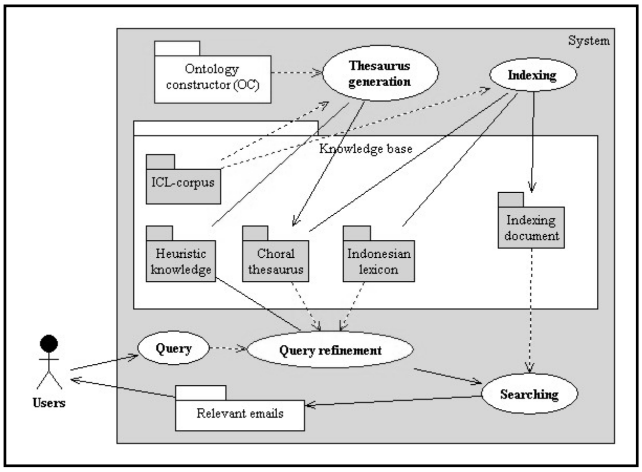

Figure 1 : Use-case diagram of text retrieval system based on ontology for Indonesian choral. (Notes: — association, → direct association, ⇢ dependency)

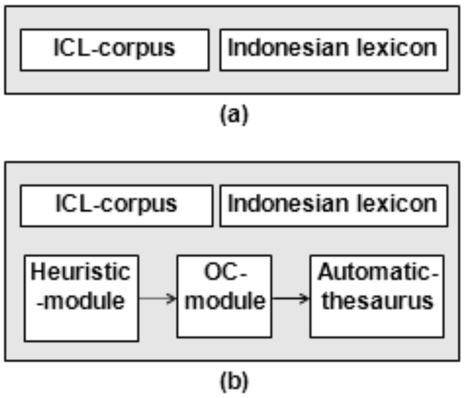

Figure 2 : The content of knowledge base in (a) empty-IR-system and (b) OC-knowledge-IR-system.

Knowledge acquisition is the main task conducted in knowledge definition stage. Some knowledge should be acquired from annotation process and the corpus, which then yields manual thesaurus, heuristic knowledge, and ontology constructor knowledge. Manual thesaurus is produced by extracting terms from ICL-corpus as well as defining their relationship manually. It is not our intention to develop a complete manual thesaurus, however this exhaustive task is essential mainly to acquire deep understanding of terms extraction and terms relation in order to construct an associative thesaurus for ICL mailing list. A number of rules suggesting heuristic tasks in order to help the ontology constructor work particularly are included in heuristic knowledge, e.g. while dealing with the natural language in email. Ontology constructor (or simply OC) is an automatic constructor of thesaurus for the ICL mailing list. The result of OC is automatically created thesaurus, or simply automatic-thesaurus, which is the choral thesaurus in Figure 1.

In knowledge definition stage, the OC is not generated yet, however the acquired knowledge on this stage is essential for the next stages of Linear Model, consecutively knowledge and system design and code and checkout. The former stage is basically design of OC, heuristic knowledge, and the Information Retrieval (IR) system into the ones ready to be encoded, hence the result on this stage will be those representation designs. Concerning the IR system, consideration of techniques in each step of the system (e.g. parser, stopwords, stemmer, indexer, searching) should be taken into account. In this respect, there is huge probability those techniques would be improved to handle the natural language of Bahasa Indonesia used in emails. Nevertheless, taking benefit from the available IR system, e.g. Indri[9],

[9] http://www.lemurproject.org/indri/

is highly possible and will save much time in order to focus on specific thing, e.g. the OC generation.

Table 1 : Summary of Linear Model implemented on the study

No.	Stages	Important Task	Main Output
1.	Planning	Corpus preparation, annotation process	Annotated ICL-corpus, TREC-like test collection
2.	Knowledge definition	Knowledge acquisition, manual thesaurus development	Heuristic knowledge, OC knowledge, Manual thesaurus
3.	Knowledge and system design	Design of OC, heuristic module, and IR system	Design of OC, design of heuristic module, and design of IR system
4.	Code and checkout	Code implementation, unification of OC and Empty-IRS	OC module, heuristic module, empty-IR-system, OC-knowledge-IR-system
5.	Knowledge verification	Technical evaluation, user's evaluation, test analysis	Test reports and analysis
6.	System evaluation	Result evaluation	Report, recommendation of revision and future study

Code and checkout stage inherently consists of code implementation of the designs (OC, heuristic knowledge, and the IR system) and then yields a an OC module, heuristic module, and an empty-IR-system. An empty-IR-system is an IR system without choral-thesaurus, OC module, and heuristic module inside it; in short an IR system without knowledge. Here, Indonesian lexicon is not considered as the system knowledge because it is ready available and become the intension of replacement with the complete IWN in the future. The code and checkout stage includes the unification of OC module and heuristic module into the empty-IR-system, thus it is called OC-knowledge-IR-system.

When the IR systems are ready, knowledge verification stage and system evaluation stage can be started. The former is devoted to the thesaurus in order to check whether it satisfies the specification requirements, while the later is to summarize what has been learned with recommendations for improvements and corrections. Section 5.3 provides the details.

Figure 3 summarizes the process and its results related with ICL corpus. The rectangle represents a process while the circle represents the result of a process. Figure 2 shows the difference between empty-IR-System and OC-knowledge-IR-system, which are in contain of its knowledge base. The dashed arrow in picture (b) of Figure 2 means dependency. It tries to show that the OC-module depends on heuristic knowledge while generating the automatic-thesaurus.

ONTOLOGY CONSTRUCTOR AND RECENT FINDINGS

A number of studies on automatic thesaurus constructor used news documents for the corpus as in [23, 27, and 28], whereas this study uses emails in ICL mailing list. Figure 4 is an example of email in the corpus. This email tells us that the writer had technical problem while posting a message and asked forgiveness to others because of receiving the same mail for three times. It is indeed has no relation with choral specifically, yet we can say that this email is not a spam. In fact, there is no single word related with choral in main message, however it is confirmed by the experts working on annotation process that none of them assigned this email as a spam. Both experts assigned "ICL" as its topic, with additional note 'masalah teknis milis' (technical problem of mailing list) given by the first expert. Although "trouble" has similarity meaning with "masalah" (problem)[10], it is more interesting that we cannot find words "ICL" or "masalah" (problem) or "teknis" (technical) or "milis" (abbreviation of mailing list) or combination of them expressed precisely on the message.

From Figure 4 we can see that a topic assigned is not described by words written explicitly in the email. Other message in corpus (let us call it Example 2) even consists of single word, "Ikutan!" (Indonesian slang word of "join"), where the other extreme one (let us call it Example 3) has no word in its main body but a sentence "wanna join" expressed on its Subject field. The first expert assigned both emails as spam because they do not have any relation with choir, while the second expert assigned nothing, which means that they are not spam. Despite the contradiction between those experts and whether those are spam or not, we should know when we read it that actual meanings of these emails are the writers wanted to join ICL mailing list.

An example of query is "saya ingin tahu tentang teknik bernyanyi yang baik" (I want to know about good singing technique). This query seems to be posted by a member who needs practical explanation about vocal technique, therefore an email about practical technique to reach high note (let us call it Example 4) may be more preferred than emails about an event of vocal workshop (let us call it Example 5) or emails about differentiation of diaphragm breathing technique (let us call it Example 6). These emails (Example 5 and 6) may be preferred while he/she is interested in deeper manner. Notice that Example 4 have no words of "teknik vokal" (vocal technique) written explicitly, but indeed in Example 5 and 6.

[10] http://thesaurus.reference.com/browse/trouble

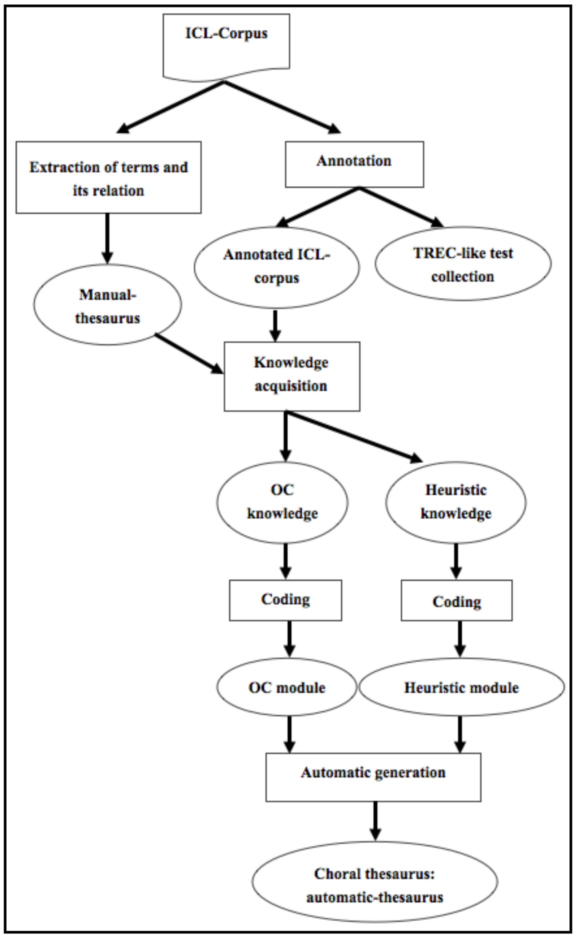

Figure 3 : Summary of process and its results related with ICL-corpus

Date: Mon, 19 Mar 2001 08:55:07 +0700
Subject: Ups sorry !
From: "Rizal A. Tandrio" <rizal74@...>

Wah ada yang enggak beres nih sama Netscape saya, sorry nih kalau sampai terima email 3 kali bukan masuknya mau kirin junk mail loh ! Cuman waktu saya tekan tombol sent unsent massage eh malah dikasih tau enggak bis kirim karena ada troble ya udah saya tekan sekali lagi, eh masih tetap sama, terakhir ya saya sent secara manual saja satu persatu, eh ternyata email yang tadi dibilang tidak bisa dikirim malah sudah terikirm semua, tu la lit nih ! Maap yaaaaaaaaaa ..... =)

Translation of the main message:
There's something wrong with my Netscape, sorry for receiving 3 emails [,] didn't intend to send a junk mail! It just when I pressed the sent unsent massage button there's a message telling me that the message was not delivered because a troble [trouble] so I pressed [the button] again, but still [received] the same [respond], finally I sent it manually one by one, and it's said that the undelivered emails were all have been delivered, so stupid! Sorrryyy...=)

Figure 4 : **An example of original email and its translation. Words in bracelet [] are editorial addition**

Those examples shows that topic annotation is beyond of written terms or terms frequency and more about semantic analysis. This is the reason that knowledge is important for a system to find out the genuine meaning of a text.

Exploration on annotation process, ICL-corpus, and manual-thesaurus development play important role in order to grasp how a topic can be assigned to an email by the experts and what kind of thesaurus should be structured. So far, it seems that related words are important to describe the semantic relations, and it confirms that an association thesaurus should benefit the intended IR system. Despite the difficulty of determining term relations automatically, it is not clear yet that relation differentiation (broader term, narrower term, and related term) is important in this particular case. Moreover, a number of studies showed promising result from developing a thesaurus using single type of association relation, e.g. based on term co-occurrence [23, 27, 28]. Simple or compound nouns proved to contribute the most in improving retrieval performance then adjectives, adverbs, and verbs in [23], but Example 2 only consists of single verb and Example 3 suggest us to see its Subject in order to understand its meaning. Further deep analysis should reveal other knowledge.

The thesaurus, either manual or automatic, may not consist of choral terms only and be complicated. Some complexity of natural language in email can be seen in Figure 4: the usage of slang word, foreign language (e.g. English), incomplete word, missing word, missing punctuation, inappropriate word, inappropriate punctuation, and special character. That email even does not have terms related with choral. Other emails may use abbreviations and different Indonesian ethnic languages (e.g. Javanese).

Meanwhile, a number of methods are being observed in order to be used in OC generation. Semi-supervised method (e.g. co-training) tempts to be considered because of the growing corpus characteristic in mailing list and relatively small number of labeled data needed. Unsupervised method (e.g. K-means method) should give much benefit with regard to unlabeled data needed which seems to be significantly easier to come in many research settings hence may be more useful in large area. A kind of certainty factor value as term weight generated automatically is inspiring to be employed as well, remember that topic assignment seems beyond written term or term frequency. However, acquired knowledge in knowledge definition stage will become prime guidance to decide the method(s) will be used and how the OC should be designed.

KNOWLEDGE VERIFICATION AND SYSTEM EVALUATION

Evaluation plays important role in order to measure OC performance. One way to determine the effectiveness of a tool is by measuring its performance based on its creation. Therefore, once we have an effective thesaurus, we can say that we have an effective OC.

There will be two kinds of formal test: technical evaluation and user's evaluation. The technical evaluation basically is performance evaluation of IR system based on recall and precision [29] whereas user's evaluation is a qualitative evaluation based on observation and written documents.

The thesaurus intends to increase the performance of retrieval system, hence the well-known measures, recall and precision, are going to be calculated, as well as the harmonic mean F [29] of recall and precision relative to the 10^{th} document in the ranking. We can see how far the thesaurus will take the progress by comparing the F values of two different systems: empty-IR-system and OC-knowledge-IR-system. Determination of the maximum value of F can be interpreted as an attempt to find the best possible compromise between recall and precision [29]. Thus, when we have a better F value on OC-knowledge-IR-system, briefly it indicates that the system has better performance because of an effective OC.

A list of topic of email in ICL-corpus which is available from the knowledge definition stage becomes a list of topic of question here. A natural form of question will be generated for each topic on that list and fed into the system as the query.

In deeper manner, the ontology evaluation will be conducted which has an objective to determine what the ontology defines correctly, what it does not, and what it does incorrectly [3].

The following criteria are going to be used to evaluate the content of a given ontology: consistency, completeness, and conciseness. These are actually qualitative measures because they are not numerical in nature. Each measure will become the effectiveness measure of automatic-thesaurus generated by the OC, but they are not combined. Therefore, it is possible to have a consistent thesaurus but not complete and concise; or any combination of them.

Recall and precision are going to be used in user's evaluation as initial data of qualitative analysis as well as supporting data. While querying, terms resulted from query refinement will be saved. This list of terms is going to be used for deeper analysis.

Similar with the technical evaluation, a question will be generated from the list of topics, however it will be fed into the system in a number of different queries yielding a number of result sets for each topic of question. By evaluating the result sets for each question topic, we can see whether the thesaurus is consistent or not. It is said to be consistent while there are no contradictory document(s) retrieved in the aggregation of the result sets.

For example, a query "saya ingin tahu tentang teknik bernyanyi yang baik" (I want to know about good singing technique) can be rephrased into "bagaimana cara bernyanyi yang baik?" (how to sing well?), or "saya butuh penjelasan tentang teknik vokal yang benar" (I need explanation about the right vocal technique). Those queries should have similar result sets because they have similar topic of question, which is 'vocal technique'. It means we should not have contradiction document(s) in any result sets of those queries, such as email in Figure 4.

Precision is fraction of retrieved documents that are relevant where a relevant document can be also seen as a document that is not contradictory with the query. In this respect, the precision value is proposed to be used as a value of consistency. For this, there will be two kinds of consistency values, for each topic of question and for the whole topics of question.

The completeness of thesaurus can be derived from observing recall value. Recall is fraction of relevant documents that is retrieved. It means, while all relevant documents in corpus is retrieved (100% value of recall) then the thesaurus is perfectly complete. As the consistency's test, this test also employ a number of queries for each topic on the topic list, and the recall value is going to be computed from aggregation of result sets of each topic as well as the whole topics. Therefore, similar with consistency test, there will be also two kind of completeness value, for each topic of question and for the entire topics of question.

There are two things should be considered in conciseness test: (a) the thesaurus should not store any unnecessary or useless terms and (b) explicit redundancies between terms do not exist. Direct inspection can expose explicit redundancy between terms either in manual-thesaurus or automatic-thesaurus. Agglomeration of terms generated in query refinement can be compared with automatic-thesaurus and also manual-thesaurus in order to recognize the unused terms. These unused terms then should be overlooked. Furthermore, we can take benefit from consistency test together with completeness test but put the focus on the irrelevant documents. Examine the irrelevant documents will reveal the unnecessary terms or unnecessary relations for specific topic. These unnecessary terms or relations should also be overlooked.

Feeding the automatic-thesaurus previously generated into the system with different corpus inside (e.g. choral-web-page corpus, music-web-page corpus) is essential in order to evaluate the sharable feature of ontology generated. This technical evaluation will use the recall, precision, and F value in order to evaluate the performance of the IR system using this particular ontology.

SUMMARY AND FUTURE STUDIES

A study focused in automatic ontology constructor is proposed. The idea came from membership experience of ICL mailing list for years. The 3,000 ICL-emails become the corpus of text retrieval system being developed and Linear Model that is known in Expert System field is applied as the system life cycle.

This study is challenging because it uses emails as the corpus and cognitive approach as the heart of the process. In order to achieve deep linguistic analysis, the natural language in emails as well as the exhaustive cognitive approach argues to be important in the process of OC generation. Experts in choral and Bahasa Indonesia are involved along the process, particularly on knowledge definition stage of Linear Model. They are working in annotation process and manual-thesaurus development. Knowledge from the intense acquisition process will be guidance to design the OC.

In order to examine the OC effectiveness, recall, precision, and F values are going to be computed for technical evaluation. However, recall and precision are also needed in qualitative evaluation mainly as indicator data for further analysis in order to inspect the consistency, completeness, and conciseness of automatic-thesaurus generated by OC. The automatic-thesaurus will also be fed into the IR system with different corpus in order to evaluate its sharable feature. The idea behind the evaluation process is an effective ontology is resulted from an effective OC.

Numerous other studies in Bahasa Indonesia may open from this point, e.g. because of the availability of new corpus and automatic OC for Bahasa Indonesia. Dealing with natural language to carry out deep understanding in a text has its own complexity. So far, the ICL-corpus reveals only few number of it. Concerning the emails, automatic identification of spam emails or not, automatic identification of emails contains foreign language, or summarization of the emails also appealing. Development of text retrieval system for Bahasa Indonesia is interesting in diverse applications while each stages of its process (e.g. parser, stopwords, stemmer, indexer, searching) is highly appealing and demand to be extended. With regard to the IWN, this study is corresponding with the merge model in the sense that it combines the manually-extracted-words from ICL-email-corpus and words of IWN being developed. However, it is still immature and requires to be improved in all fields of topics.

ACKNOWLEDGEMENT

The authors would like to thank Faculty of Mathematics, Informatics, and Mechanics of University of Warsaw and Interdisciplinary Center for Mathematical and Computational Modeling (ICM) of University of Warsaw for providing the authors a good environment and facilities of working. This study would not have been possible unless Agastya Rama Listya and Kristoforus Kuntarahadi have high commitment to be involved as the choral experts. This undergoing research is funded by European Union Erasmus Mundus "External Cooperation Window" EMMA; Specific Grant Agreement Number-2008- 4950/001-001-MUN-EWC and Duta Wacana Christian University research grant.

REFERENCES

1) Alwi, H., Sardjowidjojo, S., Lapoliwa, H., & Moeliono, A. (2003). Tata Bahasa Baku Bahasa Indonesia Edisi Ketiga. Jakarta: Pusat Bahasa dan Balai Pustaka.
2) Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition. 5, 2 (June 1993), 199-220. DOI= 10.1006/knac.1993.1008.
3) Gómez-Pérez, A., Fernández-López, M., & Corcho, O. (2004). Ontological Engineering. London: Springer-Verlag.
4) Lassila, O. & McGuinness, D. (2001). The Role of Frame-Based Representation on the Semantic Web. Technical Report KSL-01-02. Knowledge System Laboratory, Stanford University, Stanford, California.
5) Burke, R. D., Hammond, K. J., & Cooper, E. (1995). Knowledge-Based Information Retrieval from Semi-Structured Text. In AAAI Workshop on Internet-based Information System. AAAI, 15-19.
6) Ismail, M. A., Yaacob, M., Kareem, S. A., & Halim, A. H. A. (2007). Semantic Search Engine in Institutional Repository: An Ontological Approach. In Building an Information Society for All: Proceedings of the International Conference on Libraries, Information and Society (Petaling Jaya, Malaysia, June 26-27, 2007). ICoLIS 2007. Kuala Lumpur, 55-63.
7) Tang, B. & Hodges, J. (2000). Knowledge Representation, Learning, and Reasoning in WebDoc - A Web Document Classification System. In Proceedings of the AAAI-2000 Workshop on Artificial Intelligence for Web Search (Austin, Texas, July 30, 2000).
8) Brewster, C. & Wilks, Y. (2004). Ontologies, Taxonomies, Thesauri: Learning from Texts. In Proceedings of the The Use of Computational Linguistic in the Extraction of Keyword Information from Digital Library Content Workshop (Kings College, London, UK, February 05-06, 2004).
9) Noy, N. F. & Klein, M. (2002). Ontology Evolution: Not the Same as Schema Evolution. Knowledge and Information Systems. 6, 4 (July 2002), 428-440. DOI= 10.1007/s10115-003-0137-2.
10) Buitelaar, P., Olejnik, D., & Sintek, M. (2004). A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis. In Proceedings of the 1st European Semantic Web Symposium (Heraklion, Crete, Greece, May 10-12, 2004). ESWS 2004. Springer, 31-44.
11) Clerkin, P., Cunningham, P., & Hayes, C. (2002). Ontology Discovery for the Semantic Web Using Hierarchical Clustering. Computer Science Technical Reports TCD-CS-2002-25. Trinity College Dublin, Department of Computer Science, 12.
12) Adriani, M. & Manurung, R. (2008). A Survey of Bahasa Indonesia NLP Research Conducted at the University of Indonesia. In Proceedings of the 2nd International MALINDO Workshop (Cyberjaya, Malaysia, June 12-13, 2008).
13) Pisceldo, F., Mahendra, R., Manurung, R., & Arka, I W. (2008). A Two-Level Morphological Analyzer for the Indonesian Language. In Proceedings of the 2008 Australasian Language Technology Association Workshop (Hobart, Australia, December 08-10, 2008), 142-150.

14) Hartono, H. (2002). Pengembangan Pengurai Morfologi untuk Bahasa Indonesia dengan Model Morfologi Dua Tingkat Berbasiskan PC-Kimmo. Undergraduate Thesis. Call number: SK-0516, Faculty of Computer Science, University of Indonesia.
15) Adriani, M., Asian, J., Nazief, B., Tahaghogi, S. M. M., & Williams, H. E. (2007). Stemming Indonesian: A Confix-Stripping Approach. ACM Transaction on Asian Language Information Processing. 6, 4 (December 2007). DOI= 10.1145/1316457.1316459. http://doi.acm.org/10.1145/1316457.1316459.
16) Asian, J. (2007). Effective Techniques for Indonesian Text Retrieval. Doctor of Philosophy Thesis. School of Computer Science and Information Technology, RMIT University.
17) Nazief, B. A. A. & Adriani, M. (1996). Confix-Stripping: Approach to Stemming Algorithm for Bahasa Indonesia. Internal Publication. Faculty of Computer Science, University of Indonesia.
18) Putra, D. D., Arfan, A., & Manurung, R. (2008). Building an Indonesian WordNet. In Proceedings of the 2nd International MALINDO Workshop (Cyberjaya, Malaysia, June 12-13, 2008).
19) Fellbaum, C. (Ed.). (1998). WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.
20) Herdiyeni, Y. & Hasibuan, Z. A. (2003). Information Retrieval System in Bahasa Indonesia Using Latent Semantic Index and Semi-Discreate Decomposition. In Proceedings of The Fifth International Conference on Information Integration and Web-Based Applications Services (Jakarta, Indonesia, September 15-17, 2003). iiWAS2003. Jakarta, 295-304.
21) Pribadi, A. W. & Hasibuan, Z. A. (2003). Implementing Inference Network for Information Retrieval System in Indonesian Language. In Proceedings of The Fifth International Conference on Information Integration and Web-Based Applications Services (Jakarta, Indonesia, September 15-17, 2003). iiWAS2003. Jakarta, 313-322.
22) Lestari, D. P. & Furui, S. (2009). Proper Noun Adaptation for Improving a Spoken Query-based Indonesian Information Retrieval System. In Proceedings of International Conference on Rural Information and Communication Technology (Bandung, Indonesia, June 17-18, 2009), 366-371.
23) Jing, Y. & Croft, W. B. (1994). An Association Thesaurus for Information Retrieval. In RIAO 94 Conference Proceedings (Rockefeller University, New York City, October 11-13, 1994), 146-160.
24) Bly, B. M. & Rumelhart, D. E. (Ed.). (1999). Cognitive Science. California: Academic Press.
25) Voorhees, E. M. & Harman, D. (1997). Overview of the Ninth Text REtrieval Conference (TREC-9). In Proceedings of the Text Retrieval Conference (Gaithersburg, Maryland, 1997). NIST Special Publication 500-249, 1-14.
26) Giarratano, J. C. & Riley, G. D. (2005). Expert Systems: Principles and Programming, Fourth edition. Canada: Course Technology.
27) Kaji, H., Morimoto, Y., Aizono, T., & Yamasaki, N. (2000). Corpus-dependent Association Thesauri for Information Retrieval. In Proceedings of the 18th Conference on Computational Linguistics (Saarbrücken, Germany, 31 July - 04 August 2000), 404-410. DOI= 10.3115/990820.990879.
28) Lee, H., Lin, S., & Huang, C. (2001). Interactive Query Expansion Based on Fuzzy Association Thesaurus for Web Information Retrieval. In Proceedings of the 10th IEEE International Conference on Fuzzy Systems (Australia, December 2-5, 2001), 724-727.
29) Baeza-Yates, R. & Ribeiro-Neto, B. (1999). Modern Information Retrieval. New York: ACM Press.