Automatic Ontology Constructor for Indonesian Language

Gloria Virginia
Faculty of Mathematics, Informatics, and Mechanics
University of Warsaw
virginia@icm.edu.pl

Nguyen Hung Son
Faculty of Mathematics, Informatics, and Mechanics
University of Warsaw
son@mimuw.edu.pl

Abstract

Rich information is scattered under Indonesian Choral Lovers *(ICL) mailing list and many of its members prefer posting a query-mail to using the available search engine. A text retrieval system based on ontology is then proposed. However, considering the continual number of emails, developing an automatic ontology constructor (OC) will be the focus of the study while the retrieval system becomes an evaluation tool of the OC developed. Besides using 3,000 emails of ICL as the corpus, this study is challenging for it takes a cognitive approach as the heart of the process and employs Linear Model known in Expert System field as the system life cycle. The effectiveness of OC will be determined based on the automatic-thesaurus effectiveness. Performance measure of information retrieval is going to be computed in technical evaluation while qualitative measures of ontology (consistency, completeness, and conciseness) are going to be used in user's evaluation.*

1. Background

There are more than 11,000 emails in Indonesian Choral Lovers (ICL) mailing list since it was made in March 2000. This Yahoo! Groups is rich in information coming from its members that is now up to a thousand including people considered as choral experts, classical singers, and musicians in Indonesia. However, it is observed that the information is buried under the emails and hard to find, although a search facility is available. Members prefer to post their question in natural language rather then using the available search engine which contributes to the repetition of questions and discussions in ICL mailing list. The repetition makes the discussions boring, retards the learning process, and decreases the possibility of a qualified answer for a question.

An ontology-based text retrieval system which supports natural language queries becomes an idea of enhancing the effectiveness and efficiency of accessing the information in ICL. However, regarding the continual increasing number of emails, an automatic ontology constructor which has the ability to deal with natural language and semantic analysis is expected.

A number of studies on automatic thesaurus constructor such as in [1, 2, 3] worked for English and used newswire as the corpus. A study for Indonesian should be designed specifically because its morphological rules is different from English (e.g. affixes in Indonesian including prefixes, suffixes, infixes, and circumfixes [4]).

Instead of newswire, this study uses the first continual 3,000 emails of ICL mailing list. The colloquial and slang words written on those emails boost complexity, as well as foreign language (e.g. English) and Indonesian ethnic language (e.g. Javanese) particularly. In order to benefit from the natural form of Indonesian language, syntactically and semantically, a cognitive approach argues to be important in processing the corpus.

The related studies are described in section 2. The proposed system is briefly explained before stating the research problem. Linear Model as the system life cycle is presented in section 5.1 followed by the recent findings and evaluation process. Contributions and future studies will end this article.

2. Related Studies

An ongoing study to develop Indonesian WordNet (IWN) [5] is an example of a significant study being done by an Indonesian research community which is among some of the considerable efforts since the mid 1990s [6]. Relevant studies about text retrieval applications in Indonesian language are identified but none in automatic ontology constructor exclusively. A study of Pribadi and Hasibuan in [7] tried to implement inference network using Indonesian news articles while a study of spoken query-based Indonesian information retrieval was presented in [8].

A number of studies show that ontology can enhance system performance. Burke, Hammond, and Cooper study of Frequently Asked Question (FAQ) finder system in [9] shows that combination techniques

978-0-7695-4191-4/10 $26.00 © 2010 IEEE
DOI 10.1109/WI-IAT.2010.122

of information retrieval and natural language processing work better than any single approach. An ontology was generated and implemented in the search engine of institutional repository of scholarly literature of Malaysia to find conceptual relations between documents and retrieve related articles in a more efficient way [10]. This study seems to work with small data that they were experimenting with. An interesting study of [11] in WebDoc (a web document classification system) used a thesaurus and certainty factor (CF) for index generation, where 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. It is promising, although they worked on a relatively small training set and without the semantic tags on noun phrase and other contextual information.

There has been an increasing study about learning or adapting ontology dynamically in recent years such as studies in [12] and [13]. In [3], Protégé plug in was used to bootstrap a domain-specific ontology from a relevant text corpus of neurology while a study in [13] tried to automatically generate class hierarchies using conceptual clustering algorithm, COBWEB, as a basic ontology represented using Resource Description Framework (RDF) schema in a Smart Radio online song recommendation application.

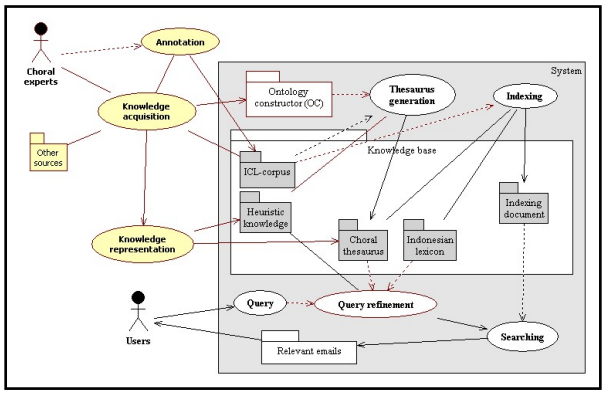

Figure 1. Use-Case Diagram of ontology based information retrieval system for Indonesian choral. (Notes: ___ association, ⟶ direct association, ----> dependency).

3. Proposed Model

Figure 1 is the use-case diagram of the proposed model for this study. The search engine which is the front-end is going to be a tool in evaluating the constructor performance. At the back-end, there is choral thesaurus as the system knowledge works for indexing and query refinement.

Experts in choral are preferred to be involved along the process, particularly on *knowledge definition* stage of the Linear Model. During annotation task, choral experts should assign topic(s) for each email and also define the highly related words written on emails with the topic given. From this task we can grasp how the topic(s) can be assigned to an email and further we can learn the relation between words and a topic semantically. Heuristic rules generated from the experts will be fed as the system knowledge that works with the constructor in order to maintain a continues updated-ontology.

There will be two kinds of information retrieval systems (IRS): empty-IRS and OC-IRS. The difference between them is basically on the knowledge base. The former will have only an Indonesian lexicon (which is already available) whereas the latest will have an Indonesian lexicon, a heuristic module and an OC module. Empty-IRS will be the baseline model of IRS in evaluation process.

4. Research Objective

The main objective of this study is to develop an automatic constructor of an associative thesaurus for a text retrieval system when emails become the corpus. Performance of the ontology constructor will be measured by its effectiveness.

5. Methodology

5.1. Linear Model

A text retrieval system based on ontology could be seen as a merge between a retrieval system and an expert system. Therefore, this study uses Linear Model that has been used in a number of expert system projects successfully as a development process [14]. It consists of six stages which are *planning*, *knowledge definition*, *knowledge (and system) design*, *code and checkout*, *knowledge verification*, and *system evaluation*. Figure 2 summarizes the task and output of each stage of Linear Model implemented during the study.

5.2. Recent Findings

Meanwhile, related words seems to be important in describing semantic relations. However, it is not clear yet whether relation differentiation (BT, NT, and NT) is necessary for this particular case; as studies in [1, 2, and 3] show promising results using a single type of association relation, e.g. based on term co-occurrence. In [1], the most contribution to improve retrieval performance were given by simple or compound nouns, but in the corpus there is an email

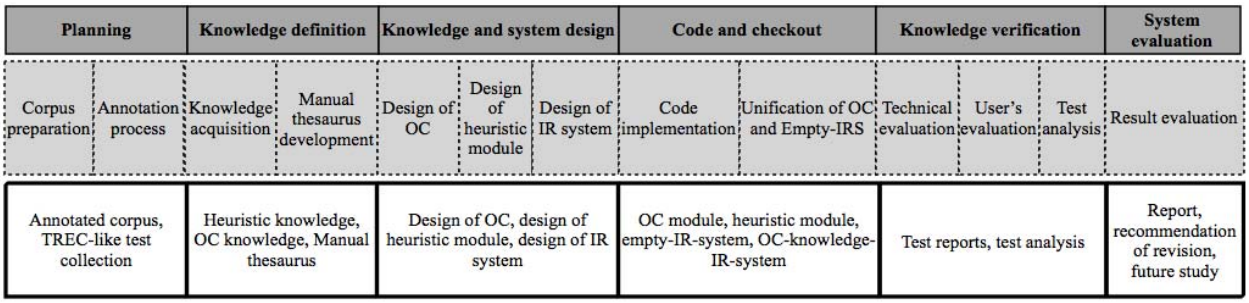

Figure 2. Summary of Linear Model process implemented in this study. First row is the stages, second row is the important tasks of each stage, and third row is the output.

which contains only single verb. An other email suggests that we read its *Subject* while the other requires us to read its forwarded message in order to understand its meaning.

Further exploration needs to be conducted as well as further consideration of a number of methods for the OC generation. Instead of the supervised method, unsupervised or semi-supervised model will be preferred because of the learning experience during annotation task, which is similar to those processes.

From 1,000 emails, it is observed that almost 50% of them have a *Subject* that describes its content. Therefore, a kind of certainty factor value as term weight generated automatically is encouraged to be implemented as well. However, above all, decision of method(s) and how the OC should be designed lays on knowledge acquired in the *knowledge definition* stage.

5.3. Evaluation

The idea behind the evaluation process is that a tool performance can be measured by the performance of its creation: thus, once we have an effective thesaurus, we can say that we have an effective OC.

Technical evaluation and user's evaluation are the formal tests of this study. Recall, precision, and the harmonic mean F [15] of recall and precision relative to the 10th document in the ranking will be calculated for technical evaluation which basically is performance evaluation of IR system. The recall and precision will also be needed mainly as initial data for qualitative analysis as well as supporting data. In this user's evaluation, three qualitative measures (consistency, completeness, and conciseness [16]) are going to be used to evaluate the content of automatic-thesaurus generated by the OC.

The *knowledge definition* stage will come up with a list of topics of email from ICL-corpus. In this stage, this list becomes a list of topic of question. A number of natural form of queries generated from each topic on the list is then fed into the system. Therefore, there will be a number of result sets for each topic of question.

The ontology is said to be consistent while there are no contradictory document(s) retrieved in the aggregation of the result sets of the retrieval system. Here, the precision value is proposed to be used as a value of consistency based on an idea that a relevant document can also be seen as a document that is not contradictory with the query. Two kinds of consistency values may be calculated: a value for each topic of question and a value for the topics of question as a whole.

Calculating the recall value from each topic of question as well as the topics of question as a whole will result in the completeness values. The ontology is perfectly complete while all relevant documents in the corpus is retrieved by the retrieval system (100% value of recall).

Instead of calculating recall and precision values, the conciseness of ontology will be evaluated through examining the irrelevant documents of consistency and completeness tests in order to recognize the redundant terms. It is a comparison between agglomeration of terms generated in query refinement and terms generated from the irrelevant documents.

The technical evaluation should be applied into the system with different corpus inside (e.g. choral webpage corpus) where the automatic-thesaurus previously generated is implemented. It is essential to investigate the sharable features of ontology generated.

6. Contribution and Future Studies

The undergoing study of IWN [5] uses expand model which is mapping the common base concepts between PrincetonWordNet (PWN) [17] synset and *Kamus Besar Bahasa Indonesia* (KBBI)[1] sense definition that is done by human annotators through

1 KBBI is a comprehensive dictionary of Bahasa Indonesia developed by *Pusat Bahasa*, the Indonesian government's centre for language development, of Ministry of Education.

web-based application. In specific manner, this study could be seen as an effort to improve and enrich the IWN. While the former study is using common base concepts, this study focuses in choral domain and crucial words in ICL emails. With this restriction the advantages are two-fold: the IWN will improve broader and deeper on particular subjects.

This study may lead to numerous other studies in natural language processing and text retrieval particularly for Indonesian language, namely developing a better parser, stopwords, stemmer, indexer, or searching algorithm.

7. Acknowledgement

This is an undergoing study under European Union Erasmus Mundus "External Cooperation Window" EMMA; Specific Grant Agreement Number-2008-4950/001-001-MUN-EWC and a research grant from Duta Wacana Christian University. It has been partially supported by grants N N516 368334 and N N516 077837 from Ministry of Science and Higher Education of the Republic of Poland.

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 hosting this research and providing us good facilities. We are grateful for the high commitment of our choral experts, Agastya Rama Listya and Kristoforus Kuntarahadi.

8. References

[1] Y. Jing and W. Bruce Croft, "An Association Thesaurus for Information Retrieval", *RIAO 94 Conference Proceedings*, Rockefeller University, New York City, 11-13 October 1994, pp. 146-160.

[2] H. Kaji, Y. Morimoto, T. Aizono, and N. Yamasaki, "Corpus-dependent Association Thesauri for Information Retrieval", *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken, Germany, 31 July - 4 August 2000.

[3] H. Lee, S. Lin, and C. Huang, "Interactive Query Expansion Based on Fuzzy Association Thesaurus for Web Information Retrieval", *Proceedings of the 10th IEEE International Conference on Fuzzy Systems*, Australia, 2-5 December 2001, vol. 3, pp. 724-727.

[4] Hasan Alwi, Soenjono Sardjowidjojo, Hans Lapoliwa, and Anton Moeliono, *Tata Bahasa Baku Bahasa Indonesia Edisi Ketiga*, Pusat Bahasa dan Balai Pustaka, Jakarta, 2003.

[5] D. D. Putra, A. Arfan, and R. Manurung, "Building an Indonesian WordNet", *Proceedings of the 2nd International MALINDO Workshop*, Cyberjaya, Malaysia, 12-13 June 2008.

[6] M. Adriani and R. Manurung, "A Survey of Bahasa Indonesia NLP Research Conducted at the University of Indonesia", *Proceedings of the 2nd International MALINDO Workshop*, Cyberjaya, Malaysia, 12-13 June 2008.

[7] A. W. Pribadi and Z. A. Hasibuan, "Implementing Inference Network for Information Retrieval System in Indonesian Language", *The Fifth International Conference on Information Integration and Web-Based Applications Services (iiWAS'2003),* Jakarta, Indonesia, 15-17 September 2003, pp. 313-322.

[8] D. P. Lestari and S. Furui, "Proper Noun Adaptation for Improving a Spoken Query-based Indonesian Information Retrieval System", *Proceedings of International Conference on Rural Information and Communication Technology(r-ICT)*, Bandung, Indonesia, 17-18 June 2009, pp.366-371.

[9] R. D. Burke, K. J. Hammond, and E. Cooper, "Knowledge-Based Information Retrieval from Semi-Structured Text", *AAAI Workshop on Internet-based Information System*, 1995.

[10] M. A. Ismail, M. Yaacob, S. Abdul Kareem, and A. H. Abdul Halim, "Semantic Search Engine in Institutional Repository: An Ontological Approach", *Proceedings of the International Conference on Libraries, information and Society (ICoLIS 2007)*, LISU, FCSIT, Kualalumpur, 2007, pp. 55-63.

[11] B. Tang and J. Hodges, "Knowledge Representation, Learning, and Reasoning in WebDoc - A Web Document Classification System", *American Association for Artificial Intelligence (AAAI) Technical Report WS-00-01*, 2000.

[12] P. Buitelaar, D. Olejnik, and M. Sintek, "A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis", *Proceedings of the 1st European Semantic Web Symposium (ESWS)*, 2004.

[13] P. Clerkin, P. Cunningham, and C. Hayes, "Ontology Discovery for the Semantic Web Using Hierarchical Clustering", *Trinity College Dublin, Department of Computer Science, TCD-CS-2002-25*, 2002, pp. 12.

[14] J. C. Giarratano and G. D. Riley, *Expert Systems: Principles and Programming*, Fourth edition, Course Technology, Canada, 2005, pp. 373-379.

[15] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, ACM Press, New York, 1999, pp. 73-81.

[16] A. Gómez-Pérez, M. Fernández-López, and O. Corcho, *Ontological Engineering*, Springer-Verlag, London, 2004.

[17] C. Fellbaum, editor "WordNet: An Electronic Lexical Database", *MIT Press,* May 1998.