Evaluation Two Label Matching Approaches
For Indonesian Language

Lintang Yuniar Banowosari, I Wayan Simri Wicaksana

Gunadarma University
Jl. Margonda Raya 100 Pondok Cina Depok 16424
(62-21) 78881112 ext.309
{lintang,iwayan}@staff.gunadarma.ac.id

Abstrak

Schema matching is a basic problem of database application, semantic web, web services in e-business, semantic-based query and etc. Today's, matching implementation is mostly done in manually. The other side, some research have designed some technique for supporting the schema matching automation in some domain. Some approaches [4], [5] have been existed for conducting the schema matching such as schema level, instance level, language based approach and so on. In this paper, we will test and investigate language based and string based method for the source which written in Indonesia language.

We will do schema matching based on terminology approach for testing sources of Indonesia language case. The reason is that most of the approach is English based. The Approach process will utilize on-line dictionary which combine with WordNet with path length method as input in language based. Matching result of our approach will be evaluated by comparing with manually result matching.

Our evaluation can give the contribution to be developed as model for mapping in interoperability on Indonesia language information source which more dynamic, semi structure and opened in Internet, as well as semantic web application or web services.

Keywords– Indonesia language, ontology, label matching, semantic similarity, WordNet.

I. INTRODUCTION

Internet has contributed great value for data exchange. On other hand, Internet introduced some new issues. Currently, information sources are more massive, distributed, dynamic and open.

One of the new problem is heterogeneity in syntax, structure and semantic. Those problems also happened in variety of information sources which use Indonesia language.

Research about semantic relatedness has been part of Artificial Intelligence and psychology for the last decade. Most of semantic relatedness relate to Natural Language Processing (NLP) which refers to Roget thesaurus. After that, WordNet became a center point for researcher to look at semantic similarity. It is because of WordNet is available for free, big, and to be designed for computer.

Most of basic characteristic of human will always try to look at relation between two concepts. For examples, all are agreed that word *automotive* has meaning *car* and *tire* has relationship with *car*, but *tree* has not. But for provide one quantitative value for showing the relation level of two concepts is difficult. As other illustration, two concepts can have relation since a concepts is more general than the other concepts (example, *car* is type of *vehicle*) or also since it is part of the other concept (example, *tire* is part of *car*).

Refer to the current solution; the approach is prepared for English language. In this paper, we will examine the usage in Indonesia language by using on-line dictionary combine with WordNet in English version. The main reason is developing WordNet in Indonesia version is very time consuming and need coordination inter-discipline. The other is utilization of on-line dictionary, it can be expanded to utilize in web services for searching and information exchange and service. In parallel, evaluation string analysis for similarity will be conducted to know the result for Indonesia language as well.

This paper have 5 sections, first section is introduction. Second section introduce schema matching, and string analysis method and WordNet. The third section is talking about representation of schema

matching in RDF/OWL, and the part 4 is design and result of the evaluation. And the last part is conclusion

II. SCHEMA MATCHING

Matching is one of important task in information interoperability. Generally, the process of matching takes as input two schemas which consist of classes, properties, rules, etc. Result of matching determines the output of relation between the elements.

The motivation of matching problem is many sources has different model to represent the content of information. For example one source has class called *Server Computer*; the other source has class called *Computer*. When a request sends to the two resources, a mapping is needed to consider relation between *Server_Computer* and *Computer*.

Many different kinds of structures can be considered as data/conceptual models: description logic terminologies, UML, XMLS, RDFS, etc. Every source can have different method to store data. Therefore, interoperability information among the sources, it is needed to match between various data/conceptual models.

Survey [5, 6] classified the schema based matching techniques from input interpretation layer as follow:
- Element level
  - Syntactic (string based, language based, constraint based)
  - External (linguistic resource, alignment reuse, upper level formal ontologies)
- Structure level
  - Syntactic (graph based, taxonomy based)
  - External (repository of structure)
  - Semantic (model based)

In our experiment, the purpose is to evaluate semantic matching for sources which written in Indonesia language. The method to test is string based and linguistic resource (thesauri).

String Analysis Method

String analysis method use the structure of strings itself (as one sequence of letter), for example find and treat *Match* and *match* is similar, but it is not always for *alignment*. Similarity name or label of classes and properties between schemas can be calculated based on the string method.

There are many ways to compare string than the way to look at string (as one sequence of the letter, one sequence of wrong string, one set of words). And the most popular is method as [**7**], compare variety string match technique, from distance function up to based on token.

Levenshtein Distance

The Levenshtein [8] is a method to calculate string distance which more sophisticated then Hamming method. It's defined for strings of arbitrary length. It counts the differences between two strings, where we would count a difference not only when strings have different characters but also when one has a character whereas the other does not. The formal definition follows. For a string s, let s(i) stand for its ith character. For two characters a and b, define

$r(a, b) = 0$ if $a = b$. Let $r(a, b) = 1$, otherwise. (1)

Assume we are given two strings s and t of length n and m, respectively. We are going to fill an (n+1)×(m+1) array d with integers such that the low right corner element d(n+1, m+1) will furnish the required values of the Levenshtein distance L(s, t).

The definition of entries of d is recursive. First set d(i, 0) = i, i = 0, 1,..., n, and d(0, j) = j, j = 0, 1, ..., m. For other pairs i, j use

$d(i, j) = \min(d(i-1, j)+1, d(i, j-1)+1, d(i-1, j-1) + r(s(i), t(j)))$ (2)

Euclidian N-Gram Distance

In mathematics, the Euclidean distance or Euclidean metric [9] is the "ordinary" distance between the two points that one would measure with a ruler, which can be proved by repeated application of the Pythagorean Theorem. By using this formula as distance, Euclidean space becomes a metric space (even a Hilbert space). Older literature refers to this metric as Pythagorean metric.

Distance between two points: [9]

$P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$, in n-space Euclidian defined as follows:

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$ (3)

2D Distance as follows: for 2 (2) points 2D:

$P = (p_x, p_y)$ and $Q = (q_x, q_y)$, distance is counted as:

$$\sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \qquad (4)$$

WordNet

WordNet is a semantic network database for English which was developed by Princenton University. Some versions of other languages also have been developed, such as EuroNet. WordNet is a lexical database which has unique meaning of word which is presented in synonym set or synset. Each synset has glossary which defines concept its representation.

For examples word *car, auto, automobile* and *motorcar* has one synset with glossary as something has four wheels vehicles which be moved by machine.

Synset is connected with variety of relation like hypernym (is type of), meronym (is part of), antonym (is opposite of) etc. If one noun A relate to noun B with "type of", so B is hypernym of A or A is hyponym of B. For example car is hypernym hatchback, or hatchback is hyponym of *car.*

Semantic similarity methods in WordNet divide into two big groups of methods, which are path length and information content.

Path length simply counting number of node or relation which connects among node in taxonomy, the shorter distance between two labels of concept, it means it has higher similarity. Path length give strength without depend on corpus statistic and it is not influenced by word distribution. But it has a weakness in taxonomy which has a uniform distance. Some examples path length approaches are Leacock-Chodorow, Resnik and Wu-Palmer. In this paper, Wu-Palmer is a model which is utilized to evaluate, its equation shown in formula 5. Wu-Palmer

$$sim_{WUP} = \max \left[\frac{2\,xdepth(LCS(a,b))}{lenght(a,b) + 2\,xdepth(LCS(a,b))} \right] \qquad (5)$$

Where length (a, b) is the number of path length between a and b, depth(LCS(a,b)) is the number of path length of general concept from a and b to root.

Information content of a node is -log of the number of all possibilities (calculation based on corpus frequency) from all words which has synset. In the other word, if p(x) is probability of one instance of x, then information content of x is -log p(x).

Representation of Schema Matching Result

RDF/OWL

In RDF, the following rule holds: anyone can say anything about anything anywhere. This means the relations between two objects may reside in several documents within the web apart from the objects. One does not have to hold the physical object to state a description about it, but only use a reference to point the object. But, this concept (as its predecessor, the concept of hyperlink) will lead to data integrity problem.

The relation, which represented as a property, is a first class object in RDF. This is different than what was done in object oriented database systems, where it is assumed that the information (properties) about an object is resided in the object.

RDF has a similarity with semantic data modeling [10] since both talk about semantics. In semantics data modeling, there is a subject that has a value (object) for a certain predicate. One different is that in RDF, properties are M-to-N relations. While in semantic data modeling, they are N-to-1 relations.

Second difference is that in RDF, a property can be a sub-property of another property. This concept has never been implemented in database, and it enables specialization of a property. Finally, RDF also allows a resource to be instances of more than one class or property. Later in this report, there is a further description about RDF.

The main theme of RDF is resource description. Therefore the base of the RDF model consists of resources. Everything is a resource or else a literal.

Anything that can be identified with a URI is a resource. This means, a resource can be a part of a document, a document, a collection of documents, or the entire Web. A resource that describes the attributes, characteristics, or inter-relations between resources, is called Property. For example, a person has age, sex, height, and weight properties. Another base element of RDF model is the triple called Statement: a resource (the subject) is linked to another resource or a literal (the object) through an arc of third resource, the predicate. A statement can be defined as: <subject> has a property <predicate> valued by <object>.

The RDF model defines a model for describing interrelationships among resources in terms of named properties and values. RDF Schema provides the mechanism for redefining the RDF model by providing an externally specified semantics to specific resources, such as: *rdfs:Class*, *rdfs:subClassOf*, r*dfs:subPropertyOf*, r*dfs:domain*, and *rdfs:range*.

Based on DAML+OIL, another language is being proposed as a standard language for the Semantic Web called Ontology Web Language (OWL). OWL is expected to meet the requirements for the web ontology language [8], such that it goes beyond RDF and RDF Schema basic primitives. OWL has several design model objectives, which are: shared ontologies, ontology evolution, and ontology inconsistency detection, balance of expressively and scalability, ease of use, XML syntax, and internationalization. The language is expected to implement six use cases: web portals, multimedia collections, corporate web site management, design

documentation, intelligent agents, and ubiquitous computing.

The ontologies and inference engines can be used to enable the current search engines to find documents with more accurate content and to make an automatic web directory (content classification). Ontologies can be embedded in image, audio, and video data to enable content searching and indexing. As machines are enabled to understand and manipulate the information in the documents, it means that the machine will become an intelligent agent. These small smart agents perform certain services. They run independently in devices within the network and communicate with each other.

This kind of technology is the basis technology for ubiquitous technology. The W3C working group, called Web- Ontology [12], is now working on the development of the language.

Representation of Mapping

RDF/OWL is one of representation 'language' for knowledge of information sources. However, standard language of RDF/OWL has limitation to represent result of matching as a mapping.

There are some possibility results of matching in logical level:

- similar relation, such as ZIP similar as Post_Code.
- super relation, such as computer super of laptop.
- sub relation, such as laptop sub of computer
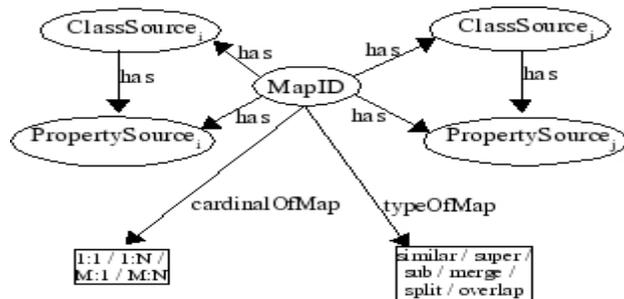- merge relation, such as Internet merge of computer and network.

Figure 1 Representation of Mapping

- split relation, such as track and line split of road network.
- overlapping relation, such as mini computer overlapping of personal computer.
- no relation, such as computer no relation with road network.

To represent result of mapping, RDF/OWL can be utilized in special model such as figure 1.

III. EVALUATION

Design of Evaluation

The goal of experiment is to evaluate matching by using string based and WordNet without develop WordNet in Indonesia version.

String based similarity will be conducted using Lavensthein (Edit Distance) and Euclidean approach. The labels which calculated are original from Indonesia words and result of translation Indonesian to English words.

For the WordNet, steps of experiment as follow:

- Select available on-line dictionary which will be used
- Select reference dictionary from good publisher.
- Calculate similarity of label based on WordNet by using label which translated from Indonesia to English. The experiment will be evaluated by consider value of Recall, Precession, and F-measure.

Figure 2 explains procedure to select appropriate on-line dictionary:

- Find the availability of on-line dictionary in Internet using search engine. In our experiment, we select 10 on-line dictionaries.
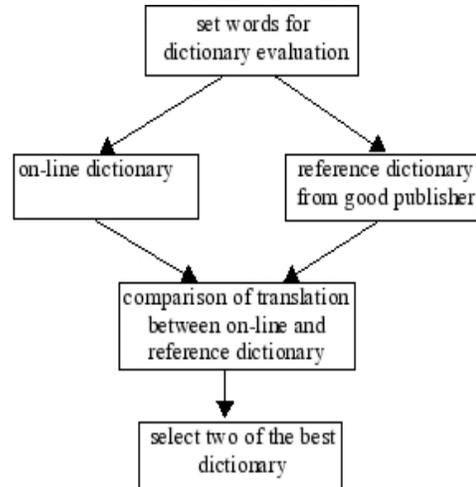

Figure 2 On-line Dictionary Selection Process

- Testing of on-line directories by using set of words (30 words).
- Translation result of on-line dictionary will be compared to the Indonesia - English dictionary by John M. Echols and Hassan Sadily which published by Gramedia. From this testing result, we will select 2 on-line dictionaries to be utilized.
- Using a tool to calculate the similarity based the above method. In this experiment use available on-line tool, for WordNet use http://marimba.d.umn.edu/ cgibin/similarity.cgi

4

Computation of similarity evaluation for Indonesia language will follow the process as depicted in figure 3, as follows:

- Compute all combination between 2 sources in one domain based on Wu-Palmer
- Filter the counting result by giving threshold value, its objective is to make easier in analysis. Assignment the threshold value conducted in try-error way to get the optimum value, start with initial value 0.7 to 1.0. This is called as counting result table (∂)
- Expert or manual matching result is β, and matching experiment result is ∂. Positive true result with this reference; it is called Δ.
- Then compute the Recall (Recall = (Δ/β)), Precession (Precession = (Δ/∂)) and F-Measure (F-measure = 2/ ((1/Recall) + (1/Precession))).
- The computing result will be shown in a table to be evaluated whether that approach is appropriate.
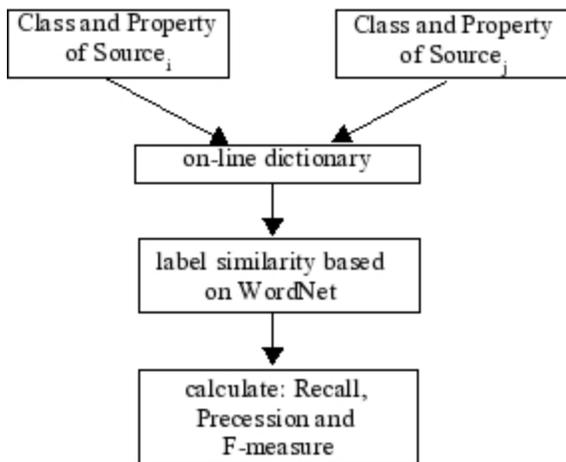

Figure 3 Indonesia Language Schema Matching Evaluation

Result and Discussion

The evaluation results are as follows:

- English translated from 2 domains, table 1 and table 2. On-line dictionary which used are
  - (i) http://www.kamus-online.com/index.php?lang=en and
  - (ii) http://www.seasite.niu.edu/Indonesian/Tata Bahasa/dictionary/Default.htm.

  For car domain is taken from http://www.mobilbabe.com/ and http://www.autocybercenter.com/. For computer domain is taken from http://www.bhinneka.com/ and http://www.belikomputer.com/

- Calculation using WordNet with Wu-Palmer formula. WordNet processing is only use output from on-line dictionary 1, since output from on-line dictionary 2 is not appropriate (table 1 and 2). Result

from WordNet table is compared with the expert result for calculate Recall, Precession and F-Measure, it can be seen in table 3.

Table 1 Translated Result for Car Domain

Source	Class	Dictionary 1	Dictionary 2
mobil-	tahun	year	year
babe	merk	tademark	NA
	deskripsi	description	NA
	harga	price	price
	kilometer	kilometer	kilometer
	kota	city	city
	foto	photo	picture
autocyber-	merk	tademark	NA
center	tahun	year	year
	warna	color	color
	lokasi	location	NA
	harga	price	price

Table 2 Translated Results for Computer Domain

Sources	Class	Dictionary 1	Dictionary 2
bhinneka	deskripsi	description	NA
	mata uang	currency	NA
	harga	price	price
beli-	produk	product	product
komputer	harga	price	price

- On-line dictionary have limited words compare to publish dictionary. From 10 on-line dictionaries which are evaluated from Internet, actually they are only 2 which are passing the evaluation by using set of words. Moreover, we only utilize one dictionary from two dictionary, because limit of words for car and computer e-business domain.
- Schema matching result in label matching terminology by using on-line dictionary and WordNet v2.1 (English language) have value is higher than minimum value of reference [1]. The other think, WordNet can perform better result compare to string based similarity.

Table 3 Recall, Precession and F-Measure

Type Exp	Recall	Precession	F-measure
WN-1	75.00%	100.00%	85.71%
WN-2	50.00%	100.00%	66.67%
Edt-1a	60.00%	75.00%	66.67%
Edt-1b	80.00%	11.43%	20.00%

Edt-2a	50.00%	100.00%	66.67%
Edt-2b	100.00%	33.33%	50.00%
Ng-1a	100.00%	16.67%	28.57%
Ng-1b	100.00%	16.67%	28.57%
Ng-2a	50.00%	25.00%	33.33%
Ng-2b	50.00%	25.00%	33.33%

WN-1: WordNet in Car e-business
WN-2: WordNet in Computer e-business
Edt-1a: Levenshtein in Car e-business (English)
Edt-1b: Levenshtein in Car e-business (Indonesia)
Edt-2a: Levenshtein in Computer e-business (English)
Edt-2b: Levenshtein in Computer e-business (Indonesia)
Ng-1a: Euclidian/n-gram in Car e-business (English)
Ng-1b: Euclidian/n-gram in Car e-business (Indonesia)
Ng-2a: Euclidian/n-gram in Computer e-business (English)
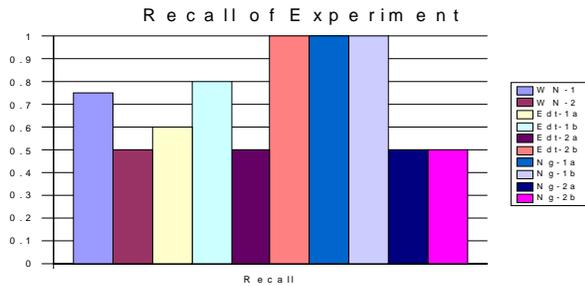Ng-2b: Euclidian/n-gram in Computer e-business (Indonesia)
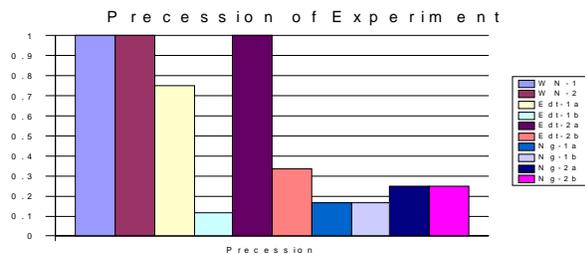

Figure 4 Recall of Experiment

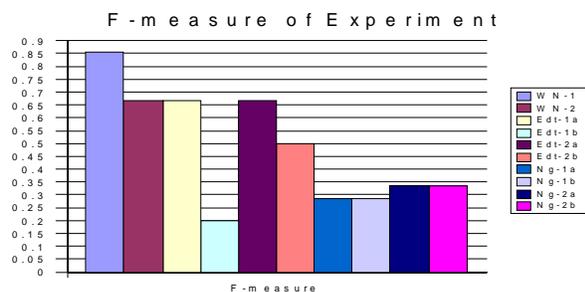Figure 5 Precession of Experiment

Figure 6 F-Measure of Experiment

IV. CONCLUSION

Combination of WordNet and on-line dictionary is acceptable method to solve diversity of element in syntactic level for Indonesia language. It is possible to bring this method to semantic diversity of sources in Indonesia language as well.

In the future we will compare this method by implementing WordNet Indonesian version and latent semantic. The main reason is many approach of matching and mapping refer to English language, but we believe in the future more and more sources of information will be in many languages.

REFERENCES

[1]. Satanjeev Banerjee and Ted Pedersen.Extended gloss overlaps as a measure of semantic relatedness. In Georg Gottlob and Toby Walsh, editors, *IJCAI*, pages 805–810. Morgan Kaufmann, 2003.
[2]. Thomas Hofmann. Probabilistic latent semantic analysis. In Kathryn B. Laskey and Henri Prade, editors, *UAI*, pages 289–296. Morgan Kaufmann, 1999.
[3]. Dekang Lin. An information-theoretic definition of similarity. In JudeW. Shavlik, editor, *ICML*, pages 296–304. Morgan Kaufmann, 1998.
[4]. Erhard Rahm and Philip A. Bernstein. A Survey of Approaches to Automatic Schema Matching. *The VLDB Journal*, 10:334–250, 2001.
[5]. Pavel Shvaiko and Jerome Euzenat. A Survey of Schema-based Matching Approaches. *J. Data Semantics IV*, pages 146–171, 2005.
[6]. Martin Warin. Using WordNet and Semantic Similarity to Disambiguate an Ontology. http://ling16.ling.su.se:8080/ PubDB/ doc repository/ warin2004usingwordnet. pdf, 2004.
[7]. Cohen, W.W., Ravikumar, P., Feinberg, Stephen, A Comparison of String Distance Metrics for Name-Matching Tasks, JCAI – AAAI Workshop, 2003.
[8]. http://www.cut-the-knot.org/do_you_know/Strings.shtml
[9]. http://www.josef-willenborg.de/java/NGram/ NGramApplet.html
[10]. J.H. ter Bekke, Semantic Data Modeling, Prentice-Hall, Hemel- Hempstead, 1992.
[11]. Resource Description Framework Vocabulary Description Language version 1.0: RDF Schema. W3C. 2002. http://www.w3.org/TR/rdf-schema
[12]. Web-Ontology Working Group. W3C, 2001, http://www.w3.org/2001/SemanticWeb/WebOnt