TELKOMNIKA, Vol.15, No.1, March 2017, pp. 407~414
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v15i1.3176

B-BabelNet: Business-Specific Lexical Database for Improving Semantic Analysis of Business Process Models

Endang Wahyu Pamungkas*[1] Riyanarto Sarno[2], Abdul Munif[3]
[1]Informatics Department, Faculty of Communication and Informatics, Universitas Muhammadiyah Surakarta, Surakarta 57102, Indonesia
[2,3]Informatics Department, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia
Corresponding author, e-mail: ewp123@ums.ac.id*[1], riyanarto@if.its.ac.id[2], munif@if.its.ac.id[3]

Abstract

Similarity calculation between business process models has an important role in managing repository of business process model. One of its uses is to facilitate the searching process of models in the repository. Business process similarity is closely related to semantic string similarity. Semantic string similarity is usually performed by utilizing a lexical database such as WordNet to find the semantic meaning of the word. The activity name of the business process uses terms that specifically related to the business field. However, most of the terms in business domain are not available in WordNet. This case would decrease the semantic analysis quality of business process model. Therefore, this study would try to improve semantic analysis of business process model. We present a new lexical database called B-BabelNet. B-BabelNet is a lexical database built by using the same method in BabelNet. We attempt to map the Wikipedia page to WordNet database but only focus on the word related to the domain of business. Also, to enrich the vocabulary in the business domain, we also use terms in the business-specific online dictionary (businessdictionary.com). We utilize this database to do word sense disambiguation process on business process model activity's terms. The result from this study shows that the database can increase the accuracy of the word sense disambiguation process especially in particular terms related to the business and industrial domains.

Keywords*: similarity, semantic string similarity, lexical database, business domain*

1. Introduction

Semantic aspect becomes a very interesting and important topic to be discussed in Business Process Management (BPM) and also Process Mining since the emergence of the semantic business process models [1, 2]. The business process activity's name often uses specific terms of business domain. Semantic aspect is used to identify synonyms, homonyms, meronymy, holonymy or different levels of abstraction in the name of business process elements [3]. The use of synonym, homonym, hypernym, and the others as a semantic aspect is to help in the searching process and enhance interconnectivity also interoperability [3-5]. Besides that, semantic aspect also helps the calculation of business process models similarity [6]. The searching process or often called business process query can also be done by considering the semantics aspect. Some business processes could have the same function but with different names. In this case, semantics can be used to search for words that have the same meaning. So that, a matching business process can be found even though the keywords is not structurally identical [4]. It uses semantic string similarity to obtain semantic word similarity [7]. Semantics itself can play a role in the lifecycle of a business process as described in [8]. Broader, semantic also can help to analyze non-functional requirement of software standard [9].

There are many lexical databases that can be used to perform semantics analysis. WordNet [10] is the one and widely used lexical knowledge for research purpose. There have also been developed a lexical database to improve the ability of WordNet, called BabelNet [11]. The emergence of BabelNet construction method was motivated by the high effort that should be done to build a new lexical database. Therefore, they tried to construct a lexical database

Received November 15, 2016; Revised December 28, 2016; Accepted January 19, 2017

automatically without had to collect the words manually. BabelNet combines lexical databases that already exists, namely WordNet with the biggest online encyclopedia Wikipedia. This method provides an algorithm to map all page on Wikipedia to the WordNet semantic network. It has the potential to cover the lack of certain vocabulary in WordNet. It also has been quite a lot of research aim to establish a new semantic network. There is some semantic network that is built based on WordNet. There is EuroWordNet which is a multilingual lexical database for several European languages [12]. The EuroWordNet structure is similar to the WordNet database. There is MultiWordNet which also developed based on WordNet. It contains information about linkages between English and Italian [13]. There are many more lexical databases which focus on languages other than English. They are BalkaNet [14], Arabic WordNet [15], and Japanese WordNet [16]. But most of them are built based on the WordNet structure.

However as far as our knowledge, until now still not found the lexical database that was built for a particular domain. Especially for term in business and industry field that often used for naming the business process activities. So it is still difficult to achieve better performance of semantic analysis on the business process model. Based on these problems, this paper presents an approach to improve the semantic analysis quality on the business process management. We build a new semantic network called B-BabelNet. B-BabelNet is created by using the same method in [11], by mapping Wikipedia page to WordNet. However, B-BabelNet will only focus on the lemmas that are related to the business domain. We also add a businessdictionary.com page as a new source to enrich the vocabulary. The semantic network will be tested to perform word sense disambiguation especially to the terms related to the business area.

2. Proposed Method

The semantic network construction method is similar to the method proposed in the development of BabelNet. However, there are a few modifications due to our scope in the business domain. The first process is the extraction of Wikipedia page focusses on the business domain. Then, we also use a business-specific online dictionary (businessdictionary.com) as additional terms to complete the vocabulary. Both of these sources are used in the metadata construction. This metadata contains the knowledge resource for the mapping process. The result of mapping process is a list of synset, both of new and existing synset. After that, we need to construct the relation between synset. The last step is indexing the synset and its relations to produce the lexical database. The flow of our proposed method can be seen at Figure 1.

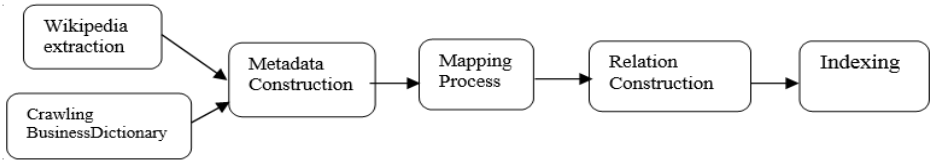

Figure 1. Proposed Method

2.1. Wikipage Extraction

The purpose of this phase is to get all the Wikipedia pages in the category of "Business". We query all the page under the category and sub-category of "Business". We need to consider that the category in the Wikipedia is not a tree but a graph in the query processes to avoid the cycle. If we look at the Wikipedia database structure in [17], the knowledge resource is not only from one table. There are two query processes to get Wikipedia page in the business category. The first process is to do a query to get a list of each page ID. After the ID is obtained, again we perform query for each page on the list to get more detail information. We use Wikipedia API [18] which provides direct access to Wikipedia database. Table 1 shows the example of Wikipage extraction for page “vendor”. We obtain information for each Wikipage as follows.

1. Page ID : Wikipedia page ID.
2. Page Title : Wikipedia page title.

3. Text : All content in Wikipedia page in HTML format.
4. Redirection pages : List of Wikipedia page(s) that used to forward to the Page that containing the actual information regarding some definitions.
5. Internal links : List of link(s) that is in the page.
6. Category : List of the category (ies) of the Wikipedia page. Wikipedia page can be assigned to one or more category.
7. isDisambiguation : Information whether the page is a disambiguation page.
8. isRedirection : Information whether the page is a redirection page.

Table 1. Example of Wikipage Extraction

```
<?xmlversion="1.0" encoding="ISO-8859-1"?>
<wikipage>
<id>14103070</id>
<title>Vendor</title>
<lemma>Vendor</lemma>
<sense></sense>
<gloss>In a supply chain, a vendor, or a seller, is an enterprise that contributes goods or services.</gloss>
<redirection>Vender_Commercialtraveller_Vendorscreening_Vendor (supply chain)_Commercial travellers_</redirection>
<link>Accounting software_Build to order_Build to stock_Customer_Distribution (business)_International Standard Book Number_Inventory_Manufacturing_Purchaseorder_Purchaserequisition_Qualityaudit_Retailing_Supply Chain Management_Supplychain_Supply chain management_Vendor, Arkansas_Vendor Managed Inventory_Warehouse management system_Wikipedia:Stub_Template:Business-stub_Templatetalk:Business-stub_</link>
<category>All stub articles_Businessstubs_Supply chain management terms_</category>
<isRedirect>no</isRedirect>
<isDisambiguation>no</isDisambiguation>
<source>wikipedia</source>
</wikipage>
```

2.2. Crawling Business Dictionary

BusinessDictionary.com [19] is an online business-specific dictionary. It contains the words and concepts specifically related to the business domain. Words on this online dictionary will be used to add vocabularies that may not have been accommodated either by WordNet or Wikipedia. There is no API or database can be accessed to get the information of businessdictionary.com page. We perform crawling process directly on the web page to get all of the information of each concept. Crawling process uses an open source library called JSoup [20]. The crawling process produces some information about each page. Here is the information obtained from crawling each page.

Table *2* is the crawling result example of business dictionary page.

1. Page Title : The title of the page and also as a lemma from the obtained word.
2. Text : Definition and explanation about the word.
3. Internal links : List of hyperlink(s) that is in the page.
4. Related terms : List of vocabulary (ies) that related to the word.

Table 2. Crawling Result Example of A Business Dictionary Page

```
<?xmlversion="1.0" encoding="ISO-8859-1"?>
<page>
<id>general ledger</id>
<title>general ledger</title>
<text> </text>
<related
term>journal_outstandingdeposit_journalizing_auditingevidence_controlaccount_capitalaccount_subledger_reconciliationaccount_organized_deferral type adjusting entry_</related term>
<link>repository_accounting_information_organization_summaries_financial transactions_subsidiaryledgers_accountingperiod_called_entry_provides_data_financial statements_</link>
<source>businessdictionary.com</source>
</page>
```

2.3. Metadata Construction

In this section, we will explain the metadata construction process for each vocabulary from the information that has been obtained in the previous stage. The purpose of metadata construction is to ease the mapping process. Also, it is used to homogenize the information from both sources, Wikipedia and BusinessDictionary.com. The difference and uniformity of the information from both sources can be seen in Table 3.

Table 3. Knowledge Resource Mapping Table

No	Metadata	Wikipedia	BusinessDictionary
1.	ID	Wikipedia page ID	Concept name
2.	Title	Whole title	Page Title
3.	Lemma	Page title without sense	Page Title
4.	Sense	Words within the parentheses	Empty
5.	Gloss	First sentence in the text	First sentence in the text
6.	Redirection pages	Redirection page list	Empty
7.	Links	List both in the text and infobox	List both in the text and telated term
8.	Category	Category list	Empty
9.	isRedirection	YES/NO according to the resource	No
10.	isDisambiguation	YES/NO according to the resource	No
11.	Source	Wikipedia	BusinessDictionary.com

Table 4. Metadata example

```
<?xmlversion="1.0" encoding="ISO-8859-1"?>
<wikipage>
<id>general ledger</id>
<title>general ledger</title>
<lemma>general ledger</lemma>
<sense></sense>
<text> </text>
<redirection>journal_outstandingdeposit_journalizing_auditingevidence_controlaccount_capitalaccount_subledger_reconciliationaccount_organized_deferral type adjusting entry_</redirection>
<link>repository_accounting_information_organization_summaries_financial transactions_subsidiaryledgers_accountingperiod_called_entry_provides_data_financial statements_</link>
<category></category>
<isRedirect>no</isRedirect>
<isDisambiguation>no</isDisambiguation>
<source>businessdictionary.com</source>
</wikipage>
```

This metadata is used as a knowledge resource in accordance with the requirements of BabelNet mapping algorithm. So that, the information extracted from both sources receive the same treatment in the mapping process. Metadata is built in XML format to ease the parsing process when retrieving the information. Metadata is constructed by using the same structure as XML from Wikipedia. However, adjustments are made to the information from businessdictionary.com. Table 4 shows the example of metadata of “general ledger” term.

2.4. Mapping Process

As explained before that the mapping algorithm is same as BabelNet mapping algorithm [11] as we can see in Figure 2. Knowledge resource is directly accessible via metadata that has been built before. Each metadata is going through the mapping process to choose a suitable synset. The probability calculation of synset is only using the context graph method [21]. Because based on the evaluation in BabelNet development, context graph method gives better results than the bag of words method [22]. The other mapping methods and also the methods to construct the relation are using mapping method that used in BabelNet.

Algorithm 1 Mapping Algorithm

Input: $Senses_{wiki}$, $Senses_{WN}$, $Senses_{BD}$
Output: a mapping $\mu: Senses_{Wiki}, Senses_{BD} \rightarrow Senses_{WN} \cup \{\epsilon\}$
1: **for each** $w \in Senses_{wiki}$ *or* $Senses_{BD}$
2: $\mu(w) \coloneqq \epsilon$
3: **for each** $w \in Senses_{wiki}$ *or* $Senses_{BD}$
4: **if** $|Senses_{wiki}$ *or* $Senses_{BD}(w)| = |Senses_{WN}(w)| = 1$ **then**
5: $\mu(w) \coloneqq w_n^1$
6: **for each** $w \in Senses_{wiki}$ *or* $Senses_{BD}$
7: **if** $\mu(w) = \epsilon$ **then**
8: **for each** $d \in Senses_{wiki}$ *or* $Senses_{BD}$ s.t. d redirects to w
9: **if** $\mu(d) \neq \in$ and $\mu(d)$ is in a synset of w **then**
10: $\mu(w) \coloneqq$ sense of w in synset of $\mu(d)$; **break**
11:**for each** $w \in Senses_{wiki}$ *or* $Senses_{BD}$
12: **if** $\mu(w) = \epsilon$ **then**
13: **if** no tie occurs **then**
14: $\mu(w) \coloneqq \underset{s \in Senses_{WN}(w)}{argmax} p(s|w)$
15: **return** μ

Figure 2. Pseudocode for Mapping Algorithm

1. (line 1-2) Initialize every Wikipedia or businessdictionary.com page w as unmapped or $\mu(w) = \epsilon$.

2. (line 3-5) For every page w that have monosemous lemma in Wikipedia, businessdictionary, and WordNet, so that w will mapped to that one sense.

3. (line 6-7) For every unmapped page w or $\mu(w) = \epsilon$, do this following step.

4. (line 8-10) every Wikipage or businessdictionary page d which redirect to w, that mapping sense have been found that is, d is a monosemous in all source including Wikipedia, businessdictionary and WordNet and such that it maps to a sense $\mu(d)$ in a synset S that also has a sense of w, we map w to the sense that suitable in S.

5. (line 11-14) For the rest page w, calculate the maximum probability $p(s|w)$ of every sense as in the Equation (1). The probability calculation is based on context graph [21].

$$\mu(w) = \underset{s \in Senses_{WN}(w)}{argmax} p(s|w) = \underset{s}{argmax} \frac{p(s,w)}{p(w)} = \underset{s}{argmax} p(s,w) \quad (1)$$

The results of the mapping process are in the form of new synsets then stored in text files. Each text contains synset information that later can be retrieved for another purpose. This file is indexed using Apache Lucene library [23] to make retrieval process easier. This index will greatly facilitate the process of querying particular synset. The same method is used to save the relation of each synset. Therefore, the development process B-BabelNet produces two indices, synset index or often called dictionary and relation index.

2.5. Constructing Synset Relation

The last step in the process of building the lexical database is building synset relation. The relation is given based on the knowledge resource used during the mapping process. The purpose of building synset relation is to enrich the relation so that the database has high coverability. Generally, the methods to relate the synset are same as the method used in the construction of BabelNet [11]. But in our case, it should be checked whether the knowledge resource is become a synset or not. Because it is possible that the knowledge resource is not included in the domain of business. If it happens, the knowledge resource is not formed into a synset and does not need to be related. The relation definition of each synset is saved into a text file. So one synset will have a relation file that contains the ID of its synset and all the synsets ID that related to it. Furthermore, all relation files will also be indexed using the same way in the construction of synset.

3. Experimental Result

We build a web-based application to facilitate the experiment. It is built by using Java and utilizes multiple open source libraries. JSoup is used for crawling BusinessDictionary.com

website, and JWNL library is used to access WordNet database [24]. Apache Lucene is used to perform indexing and searching process of synsets. We perform two evaluation, mapping evaluation to examine the mapping process and WSD evaluation to evaluate the process of word sense disambiguation. In the experiment, we use accuracy as result evaluation. Equation (2) is used to calculate the accuracy value.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative

3.1. Mapping Evaluation

In this section, we would explain the evaluation of the mapping result. We created a correct mapping standard of Wikipedia pages to WordNet sense. The mapping standard is created manually by experts. This list consists of 120 Wikipedia pages that mapped correctly to WordNet sense. This list is compared with the results of the mapping method used in this study. In addition, we also evaluate the mapping result of businessdictionary.com page using the same evaluation method. The data are taken randomly as many as 120 words.

Table 5 is the result of the accuracy of the businessdictionary.com page mapping process. From 120 pages used in the evaluation, we found that 93 pages are polysemous and 27 pages are monosemous. Test result shows that the mapping process of businessdictionary.com page has an accuracy of 0.858. Table 6 is the testing result of Wikipedia page mapping process. It turns out that 95 pages are monosemous and 25 pages are polysemous. The accuracy of the results is 0.891.

Table 5. BusinessDictionary Mapping Result

Mapping Word from BusinessDictionary		True	False
	True	103	0
	False	17	0

$$Accuracy = \frac{103+0}{101+0+17+0} = \frac{103}{120} = 0.858$$

Table 6. Wikipedia Mapping Result

Mapping word from Wikipedia		True	False
	True	107	0
	False	13	0

$$Accuracy = \frac{107+0}{107+0+13+0} = \frac{107}{120} = 0.891$$

Overall accuracy for the mapping process is 0.875. This is better when we compare it to BabelNet accuracy that is around 0.82. But we cannot judge that our algorithm is better. The result may be influenced by the number of datasets used in the experiment that is not as much as that of the experiments conducted in BabelNet.

3.2. Word Sense Disambiguation Evaluation

In this section, we describe the evaluation of the word sense disambiguation result. As explained before, the word sense disambiguation method is implemented by using an unsupervised graph-based method. B-BabelNet is used as the dictionary. We calculate the accuracy of word sense disambiguation process for certain words that have been determined. The result will be compared with word sense disambiguation using Babelfy. Babelfy is a word sense disambiguation system that was developed by utilizing BabelNet as lexicon [25].

In this experiment, we use business process models in PetriNet notation as a dataset. The dataset is taken from a running ERP business process, consist of 83 different activities. We perform word sense disambiguation to the activity name of the business process model. The purpose of this experiment is to show the performance of B-BabelNet in assisting the process of semantic analysis on the business process model. The result of word sense disambiguation process will be matched to a gold standard that was constructed. There are 83 activities that have been labeled with the proper sense of the gold standard. This standard is constructed by some expert that are quite familiar with the lexicon and the sense of it.

Figure 3 shows that our proposed method is able to get a slightly better acuracy than Babelfy. From the diagram above, we can see that the accuracy of Babelfy is 0.843 and our proposed method is 0.88. This happens because we use the database that can handle the business specific words.

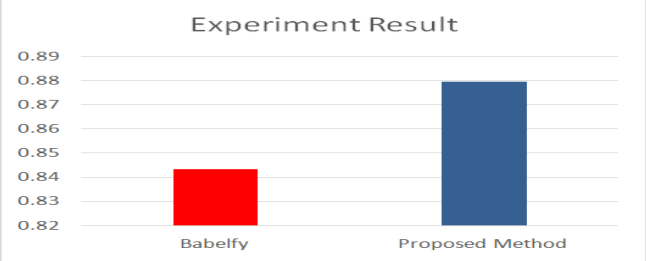

Figure 3. Experiment Result

4. Conclusion

In this study, we propose the construction of a new lexical database, called B-BabelNet. This particular lexical database contains specific concepts that related to business and industry domains. B-BabelNet is built based on the needs to improve semantic analysis of business process management. While the existing lexical database is not able to meet the needs regarding to the completeness of concepts. B-BabelNet is constructed by using the same method in BabelNet, by mapping the Wikipedia page to WordNet. In addition, we add other sources namely BusinessDictionary.com.

The result of the evaluation shows that B-BabelNet is able to provide better accuracy in the process of word sense disambiguation especially to the terms that related to the business domain. These kinds of terms are often used as the name of business process activities. A better result is obtained when we compare it to Babelfy that uses BabelNet as a lexicon. So, we can conclude that the B-BabelNet can be a solution to help semantic analysis on business process management. Furthermore, the accuracy of mapping Wikipedia page is better than mapping businessdictionary.com page. This happens because Wikipedia page does have more complete knowledge resource than the businessdictionary.com page. Generally, the result shows that the accuracy of mapping process is quite good.

However, from this research there are some things that can be improved. There is a need for improvement methods of Wikipedia database extraction. Because there are many pages found under the category of business was irrelevant. Thus, lowering the quality of the database. We also need to make improvements the mapping algorithms between Wikipedia and WordNet. Because in certain cases, the existing algorithm imposes to keep the page mapped to the same lemma. Though it is possible that the lemma of Wikipedia will enrich the generated database. Still, the database from this research has opened the opportunity to expand new research topics both in the field of natural language processing and business process management. This research will open up another opportunity to build a lexical database that related to specific field for example for words related to medical, military, etc. Surely the resource used will be depending on the needs of each field.

References

[1] M Hepp. Semantic business process management: A vision towards using semantic web services for business process management. 2005.
[2] R Sarno, WA Wibowo, K Kartini, Y Amelia, Y Rossa. Determining Process Model Using Time-Based Process Mining and Control-Flow Pattern. *TELKOMNIKA.* 2016; 14(1): 349-360.
[3] M Ehrig, A Koschmider, A Oberweis. *Measuring similarity between semantic business process models.* In Proceedings of the fourth Asia-Pacific conference on Comceptual modelling. 2007; 67.
[4] Awad A Polyvanyy, M Weske. *Semantic querying of business process models.* In EDOC'08. 12th International IEEE. 2008.
[5] R Sarno, CA Djeni, I Mukhlas, D Sunaryono. Developing a Workflow Management System for Enterprise Resource Planning. *Journal of Theoretical & Applied Information Technology.* 2015; 72(3).
[6] R Sarno. *Clustering of business process fragments.* In IC3INA. 2013.
[7] R Dijkman, M Dumas. Simil arity of business process models: Metrics and evaluation. *Information Systems.* 2011; 36: 498-516.
[8] D Mederios, A Alves. *An outlook on semantic business process mining and monitoring.* In On the Move to Meaningful Internet Systems. 2007: 1244-1255.

[9] DA Ramadhani, S Rochimah, UL Yuhana. Classification of Non-Functional Requirements Using Semantic-FSKNN Based ISO/IEC 9126. *TELKOMNIKA.* 2015; 13(4).
[10] GA Miller. WordNet: a lexical database for English. *Communication of the ACM.* 1995; 38(11): 39-41.
[11] R Navigli, SP Ponzetto. BabelNet: The automatic construction, evaluation, and application of a wide-coverage multilingual semantic network. *Artificial Intelligence.* 2012; 193.
[12] P Vossen. *Introduction to eurowordnet*. In EuroWordNet: A multilingual database with lexical semantic networks. SPRINGER, Netherlands.1998: 1-17.
[13] E Pianta. *MultiWordNet: Developing an aligned multilingual database.* In Proceedings of the 1st International global Wordnet. 2002.
[14] D Tufis, D Cristea, S Stamou. BalkaNet: Aims, methods, results and prespective. A general overview. *Romanian Journal on Science and Technology of Information.* 2004: 9-43.
[15] W Black. *Introduction the Arabic WordNet project.* In Proceedings of the 3rd International Global WordNet Conference. 2006.
[16] Isahara, Hitoshi. Development of the Japanese WordNet. 2008.
[17] WikiMedia. WikiMedia Database Schema.
[18] Wiki. API:Main Page.
[19] BusinessDictionary.
[20] Hedley, Jonathan. JSoup: Java HTML Parser.
[21] R Navigli, L Mirella. *An experimental study of graph connectivity for unsupervised word sense disambiguation.* In Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2010; 32(4).
[22] R Navigli, S Ponzetto. *BabelNet: Building a very large multilingual semantic network.* In Proceedings of the 48th Annual Meeting of the Association for Computational Lingustic. 2010.
[23] Apache Lucene Core. Apache.
[24] JWNL. Available: http://sourceforge.net/projects/jwordnet/.
[25] Moro F Cecconi, R Navigli. *Multilingual Word Sense Disambiguation and Entity Linking for Everybody.* In Proceedings of the 13th International Conference on Semantic Web. 2014.