Performance Improvement of Business Process Similarity Calculation using Word Sense Disambiguation

Endang Wahyu Pamungkas[1], Riyanarto Sarno[1], Abdul Munif[1]

***Abstract* - Similarity calculation between Business Process Models (BPM) has an important role in the process of managing BPM repository. One of its uses is to facilitate the searching process of a model in the repository. Similarity calculation between business processes is closely related with semantic string similarity. Semantic string similarity is usually performed by utilizing a lexical database, such as WordNet, to find the semantic meaning of words. The problem in WordNet is that this lexical database contains terms wich have more than one meaning or polysemous. Selecting the wrong meaning will decrease the accuracy of similarity calculation process. In this study, we will try to improve the accuracy of similarity calculation of business processes using Word Sense Disambiguation (WSD). The main purpose is to eliminate the ambiguity of polysemous words before calculating the similarity value. WSD is performed by unsupervised methods based on the value of graph connectivity. Then, we used a lexical database that is focused in the business and industry field. The results from this study is able to achieve higher accuracy of the sense selection process for terms especially terms that are related to business and industrial domains. It will also increase the accuracy of similarity value calculation between the business process models*.**

***Term Index* - Word Sense Disambiguation, string semantic similarity, business process model similarity.**

INTRODUCTION

Semantic string similarity has many applications and benefits for business process management. Generally, the semantic similarity of a text is often used to find the value of the similarity between words. Classification of documents as well as data retrieval are few of well-known technique that also utilizing this method. In business process management, This method also used in the process of clustering business process model. The semantic string similarity algorithm is used to calculate the value of similarity between activities in business process model [1]. Besides that, this technique is also used in the process of business process discovery [2]. In the discovery process, semantic aspects is used to obtain models that have similar function but with different syntax as keyword. So, user can obtain models using keywords without have to use the exact same syntax as the model notation [3].

There is a tool developed by Princeton University that already well-known and widely used to find similarity value between texts, called WordNet. However, the use of WordNet for words that are related to business processes is still facing many problem. Although it is constructed in the form of graph, this lexical database is still not able to handle the ambiguity of words. This case usually occurs in polysemous words. For example, the word "call" can be define as "command to come" or "a telephone connection". Moreover, it has 30 gloss with different meaning in WordNet. Hence, the topic of word sense disambiguation is still a concern in the field of natural language processing. This will also affect the accuracy of semantic string similarity calculation.

Based on these problems, this study presents a method to improve the accuracy of string similarity value calculation. The purpose is to eliminate the ambiguity of the word using WSD before we calculate the similarity value. WSD process will be implemented using unsupervised methods and utilized a dictionary. Selection of the correct meaning is determined by the value of connectivity graph that formed based on lexicon or dictionary. We use WordNet as dictionary and some vocabulary in business domain for addition. The results of this study are expected to improve the performance of semantic string similarity calculation and especially for the similarity between business process models.

METHOD

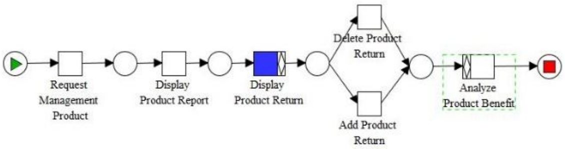

Figure 1. Business Process Model example 1.

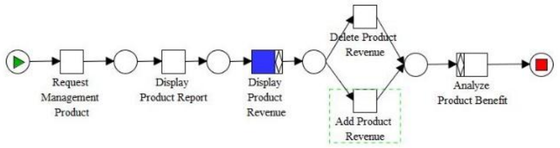

Figure 2. Business Process Model example 2

Business process similarity is highly correlated to string similarity. The label of every element in the model is compared in the process of calculating process business similarity. String similarity itself also has been developed to improve the accuracy of the comparison process. Semantic string similarity is one of the development. In semantic similarity, the similarity value is obtained not only by looking at the structure of the word but also from the meaning of the word. So, the calculation process itself requires the assistance of a lexical database. WordNet is a lexical

[1]Endang Wahyu Pamungkas, Riyanarto Sarno, and Abdul Munif are with Departement of Informatics, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Surabaya. Email : endang.wahyu.pamungkas10@mhs.if.its.ac.id, riyanarto@if.its.ac.id, munif@if.its.ac.id

database that often used to assist the calculation of semantic string similarity. Many methods have been developed to calculate the value of semantic similarity string such as Path Length, Leacock & Chodorov, and Wu Palmer. In general, these methods calculates the distance between words in WordNet and take the shortest distance as the selected meaning. The shortage is that these methods are not considering the context of words in a sentence. Thus, it is possible that the shortest distance is not necessarily the correct meaning.

Word Sense disambiguation is the process of eliminating ambiguity in the meaning of polysemous words. In this study, WSD is regarded as an essential step before calculating the value of the similarity between words. Word with the same syntax does not necessarily have the same meaning. We utilize WordNet lexicon to perform unsupervised method. The selection of word sense is determined from the inverse path length sum values [4] based on the graph structure that is formed. The graph itself is constructed from the traversal process within the structure of WordNet.

As an example, we compare the model in Figure 1 and Figure 2. We try to calculate the value of the semantic string similarity of the term "display product return" with the term "display product revenue" using the help of WordNet. Both of these terms have common initial words, "display" and "product". So, the similarity value of these words in semantic is 1. There is a difference in the last word, "return" and "revenue". By using Path Length the meaning of the word "return" is defined as "the income or profit arising from such transactions as the sale of land or other property" and the meaning of the word "revenue" is defined as "government income due to taxation". Hence without seeing the context of the word in the sentence, the similarity value is 0.25.

But, if we perform word sense disambiguation process in advance to get the meaning of a sentence based on the context, the word "return" is defined as "the income or profit arising from such transactions as the sale of land or other property" and the word "revenue" defined as "the entire amount of income before any deductions are made". The similarity value is 0.17. So that, we can conclude that the shortest distance is not necessarily provide the correct meaning of the word. We can conclude that WSD process is necessary to improve the accuracy of semantic string similarity calculation.

RESULT AND DISCUSSION

The results shows that the proposed method can achieve high accuracy even for the terms that are specific to business domain. The accuracy value is equal to 0.925. Only 9 terms from 120 terms is given the wrong sense. Therefore, we can conclude that the method is suitable to calculate the similarity value of business process implemented only one giving any sense for nine terms. Because the terms that are used for experiments is already represent the name of the business process activities.

Furthermore, we use real business process model as dataset. There are 32 business process models in Petri Net notation with a total of 69 different activities. We calculate semantic string similarity value for each activity and compare the result with Path Length method. As evaluation, we calculate accuracy value for the chosen sense. So that, we have to construct a gold standard which contains proper sense of the activities. Our proposed method can get 0.91 of accuracy. While path length method that also use WordNet get an accuracy of 0.88. So we can see that our proposed method is able to improve the semantic string similarity process and also similarity of business process. Even though, both method use WordNet as lexicon, our proposed method can give better accuracy.

CONCLUSION

The evaluation and the testing result shows that the proposed method has a good performance. Even for the terms in business domain that used in business process activities. The accuracy for SAP code dataset terms is 0.92. Besides, we also evaluate our method using dataset of real business process models. From the result, we can conclude that our proposed method can give better accuracy than Path Length method while choosing the correct sense. From these results we can concluded that the WSD method is very suitable to improve the performance of business process similarity value calculation. However, this research found some improvements that is important to be done. There are many terms in the field of business domain that have not been accommodated by WordNet. Thus, cause a decline in the accuracy of Word Sense disambiguation process. Therefore, at this time we also are currently developing a lexical data bases which specifically includes words within the business domain. Moreover, there is also a need to test another method for calculating graph connectivity because it is possible that there are other methods for calculating connectivity graph which have better accuracy.

REFERENCES

[1] R. Sarno, "Similarity of business process fragments," in *Computer, Control, Informatics, and Its Applications IC3INA*, 2013.

[2] R. Sarno, C. A. Djeni, I. Mukhlash and D. Sunaryono, "Developing A Workflow Management System for Enterprise Resource Planning," *Journal of Theoretical and Applied Information Technology,* vol. 72, no. 3, pp. 412-421, 2015.

[3] A. A. A. Polyvyanyy and M. Weske, "Semantic querying of business process models," in *Enterprise distributed object computing conference EDOC'08 12th International IEEE*, 2008.

[4] S. Brin and M. Page, "Anatomy of Large-Scale Hypertextual Web Search Engine," in *Proc. Seventh Conf. World Wide Web*, 1998.