Multi-language Dictionary Integration by Using Semantic Web

D Gunawan, R F Rahmat, F Baggio
Department of Information Technology, Universitas Sumatera Utara, Medan, Indonesia
E-mail: danigunawan@usu.ac.id

Abstract. Some languages do not have a direct translation to another language. This condition might occur to local language in the certain country. This research tackles this issue by providing an indirect translation. Indirect translation means the translation is done by linking the closest language which has a direct translation with the target language. Furthermore, this research provides integration among several dictionaries data which lie in third-party servers and all translation at once. This research yields multi-language dictionary application which utilizes semantic web technology to provide data integration. As the conclusion, The Internet speed and the availability of the third-party servers influence the performance of the multi-language dictionary application.

1. Introduction
A word in a language is usually translated to another language by using direct translation between them. However, not all languages have a direct translation to each other. Some languages are only used in the certain country. For example, Indonesia has many local languages, such as Batak language, Karo language, Javanese language, Sundanese language and many more. One cannot directly translate Batak language into English because there is no available direct translation.

The purpose of this research is to provide translation between two languages that have no direct translation. We propose the method for indirect translation. Indirect translation means a language can be translated to another language although there is no direct translation between both of them. In addition, this research also integrating several dictionary data and provide the translation from several languages at once, based on the available dictionary. We utilize semantic web technology to accomplish the purposes of this research. Semantic web technology approach is proposed in many applications including open data [1] or e-commerce protocol for data exchange [2].

This paper is organized into five sections. The first section is the introduction. It describes the background and purpose of the research. The second section discusses the previous works that related to this research. The third section is the analysis and design. It describes the architecture of the multi-language dictionary application. The fourth section discusses the implementation of multi-language dictionary application. The last section concludes the entire research.
2. Related Works
Research conducted by Zan [3] translated the Chinese language to English by using lexicon approach by merging data from two sources, i.e. Chinese Concept Dictionary (CCD) and bilingual seed set. The data extraction was conducted by using dictionary-based and seed knowledge method.

Another research conducted by Nien [4] adds the word translation by utilizing machine learning to perform hierarchical classification based on the word meaning and semantic word group. In addition, the greedy algorithm was applied to avoid the word ambiguity. This research shows hit-rate 77% and accuracy 70%.

Salam [5] built a translation machine for the Bengali language by using Example-Based Machine Translation (EBMT) and IPA-based transliteration (IPA stands for International Phonetic Alphabet). This system matched the closest semantic relation with WordNet based on input word. This result improves the quality of the translation up to 16 points.

Navigli and Ponzetto [6] implements the semantic network based on statistical machine translation to build encyclopedia dictionary called BabelNet. This research has two data sources, that is WordNet and Wikipedia to build semantic relation. To avoid word disambiguation, this research utilizes the bag-of-words method and the graph method. This research yields the accuracy of data mapping of 3 million contexts with the average of 8.6 labels per context for extraction data process. In another research [7], Qiu built the semantic relation such as hypernym, synonym, antonym and others by utilizing semantic ontology.

In 2014, the research conducted by Gratta and Nahli [8] complement Arabic WordNet by the utilizing heuristic method to supplement synonyms. The data integration is done by extracting English words by using the bilingual dictionary called AraMorph DictStems. The extracted words will be matched with others to search the synonym set. This research increases the number of synonym words in Arabic as much as 20.7%.

Bond [9] merged three WordNet dictionaries, such as WordNet (English), Malay WordNet, and Indonesian WordNet. This research utilized the merge and extend method. Extend method uses the synonym set of Princeton WordNet (PWN) and supplement lemma to target language to maintain the original structure of WordNet. This research yield the merged WordNet with the accuracy up to 78%. Another research by Zhang and Hasi [9] also built a multi-language dictionary by combining the Mongolian language, Chinese and English. They built the dictionary based on WordNet and utilized the semi-automatic matching method. However, there are some words in the Mongolian language that have no relation to Chinese language and English.

3. Analysis and Design
3.1. General Architecture
Figure 1 shows the multi-language dictionary application runs on top the Apache web server. The business logic for the web application was built by using PHP and AJAX. It utilizes JSON serialization to provide data exchange between the web application and the browser.

All of the dictionaries data are stored in the RDF format. The storage of Bahasa Indonesia and Bahasa Batak dictionaries are located in the local server. The structure of the dictionary is based on the lexical database for Bahasa Indonesia [10]. The bridge database to connect each dictionary is also stored in RDF format and located in the local server. The Apache Jena Fuseki provides the service for all the data storage in the local server. On the other hand, the WordNet dictionary is located in the separate server. The storage server of WordNet is provided by the Princeton University. The SPARQL endpoint for the WordNet is available from http://wordnet-rdf.princeton.edu/sparql/.

To obtain data that is stored in RDF storage requires SPARQL, an RDF query language. SPARQL is an acronym for SPARQL Protocol and Query Language. This is a semantic query language to retrieve and manipulate data stored in RDF format. Naturally, PHP does not
support SPARQL. Thus, the web application requires PHP SPARQL library that is available from https://github.com/cgutteridge/PHP-SPARQL-Lib in order to retrieve the RDF data. This library is the part of the query engine as shown in figure 1.

3.2. Indirect Translation
One of the problems in multi-language dictionary application is the lack of direct translation data between two languages. This research proposes the bridge to cover the lack of direct translation data by linking the available direct translation to build the relationship between those languages. Therefore, the translation will be available although there is no direct translation between them. We named this method indirect translation. As illustrated in figure 2, there is only two direct translation i.e. Bahasa Indonesia English and Bahasa Indonesia Batak language. There is no direct translation available between English and Batak language. By utilizing the bridge, the translation between those languages is possible.

3.3. Data Source
This research uses three data sources. First of all, we use WordNet that is available from https://wordnet.princeton.edu/wordnet/download/ as the English dictionary. In addition, we also utilize it to obtain the Bahasa Indonesia translation as the data source for Bahasa Indonesia dictionary. The source for translation in Batak language to Bahasa Indonesia is available from http://www.kamusbatak.com. However, the Batak language is not available in RDF format. Therefore, the source of Batak language dictionary (in comma separated version) should be converted to RDF format.
4. Implementation
The multi-language dictionary application is deployed in local computer web server. However, it requires the Internet connection to communicate with the WordNet. As the main target user of the multi-language dictionary application is Indonesian, therefore we use Bahasa Indonesia for the user interface of the application. We will provide the translation in English for the user interface.

Figure 3 illustrates the input data. Above the input data textbox is the title of the web application called "Kamus Multibahasa" or "Multi-language dictionary" in English. In this case, the input word in Bahasa Indonesia is "makan". Figure 4 shows the pop-up window after the user click "search" button (The button next to the input box in figure 3). The window consists of four parts. Each part can be accessed by clicking the icon + (plus sign) or the part name. The first part provides the definition of the input word (figure 4). As shown in figure 4, the definition part title has the green background (It has "definisi" title which means "definition" in English).

![Figure 2. Indirect translation structure](image1)

![Figure 3. Input word](image2)
Figure 4. The definition of the word "makan"

The second part provides the synonym of the input word (figure 5). Figure 5 shows the title "sinonim" which means "synonym" in English. The last two parts depend on the input word’s language. For example, if the input word is in Bahasa Indonesia, then the last two part will be the translation in English and Batak. Therefore, if the input word is in English, then the last two part will be the translation in Bahasa Indonesia and Batak. The similar condition will apply to the input word in Bahasa Batak. Figure 3 shows the input word is in Bahasa Indonesia. Then the last two parts are the translation in English (figure 6) and Bahasa Batak (figure 7). The titles of these parts are "Translasi Inggris" and "Translasi Batak" for "English translation" and "Batak Translation" respectively. Figure 6 shows that the translation of the word "makan" in Bahasa Indonesia is "eat" in English. Meanwhile, the Batak translation for the word "makan" is "tuduk" (figure 7).

As the English translation is obtained from the third-party server, then the Internet connection is required. In addition, the application will not able to obtain the English translation if the WordNet server is down. Therefore, multi-language dictionary depends on the availability of other servers.
5. Conclusion
According to the test by using the input word in Bahasa Indonesia, English and Batak language, the multi-language dictionary application obtains the definition, synonym and translations successfully. It is able to integrate several dictionaries from different data sources and provide the translation between two languages although there is no direct translation between both of them. We found that there are several factors that influence the multi-language dictionary application such as the Internet connection speed and the availability of the RDF storage. The good internet connection speed is required because the communication between the application and RDF storage depends on the Internet network. In addition, the availability of RDF storage will affect the translation result. If the RDF storage is not active, then the application will not be able to obtain the translation result. Therefore, the multi-language dictionary performance depends on the availability of the third-party servers.
References

[1] Gunawan D and Amalia A 2016 The implementation of open data in Indonesia 2016 International Conference on Data and Software Engineering (ICoDSE) pp 1–6

[2] Gunawan D 2015 Protocol for e-commerce data harvesting 2015 International Conference on Technology, Informatics, Management, Engineering Environment (TIME-E) pp 1–5

[3] Zan H, Duan G and Fan M 2007 Single word term extraction using a bilingual semantic lexicon-based approach Third International Conference on Natural Computation (ICNC 2007) vol 5 pp 451–456 ISSN 2157-9555

[4] Nien T y, Ku T, Huang C c, Chen M h and Chang J S 2009 Extending bilingual wordnet via hierarchical word translation classification Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1 URL http://aclweb.org/anthology/Y09-1040

[5] Salam K M A, Setsuo Y and Nishino T 2011 Translating unknown words using wordnet and ipa-based-transliteration 14th International Conference on Computer and Information Technology (ICCIT 2011) pp 481–486

[6] Navigli R and Ponzetto S P 2012 Artificial Intelligence 193 217 – 250 ISSN 0004-3702 URL http://www.sciencedirect.com/science/article/pii/S0004370212000793

[7] Qiu L 2013 The design and implementation of tibetan-chinese online dictionary based on semantic ontology 2013 6th International Conference on Intelligent Networks and Intelligent Systems (ICINIS) pp 256–259

[8] Gratta R D and Nahli O 2014 Enhancing arabic wordnet with the use of princeton wordnet and a bilingual dictionary 2014 Third IEEE International Colloquium in Information Science and Technology (CIST) pp 278–284 ISSN 2327-1884

[9] Bond F, Lim L T, Tang E K and Riza H 2014 NUSA: Linguistic studies of languages in and around Indonesia 57 83 – 100 URL http://hdl.handle.net/10108/79286

[10] Gunawan D and Amalia A 2017 IOP Conference Series: Materials Science and Engineering 180 012052 URL http://stacks.iop.org/1757-899X/180/i=1/a=012052