User participation in building language repository: the case of Google Translate

H Thamrin¹, G Ariyanto¹, E W Pamungkas¹ and Y Sulistyono²

¹Department of Informatics, Universitas Muhammadiyah Surakarta, Jl. A Yani Pabelan, Kartasura, Sukoharjo 57162, Indonesia
²Department of Education of Bahasa Indonesia, Universitas Muhammadiyah Surakarta, Jl. A Yani Pabelan, Kartasura, Sukoharjo 57162, Indonesia

*husni.thamrin@ums.ac.id

Abstract. A language repository is a place that stores language resources, such as dictionary or corpus. Accurate and large size repository is important to improve the performance of natural language processing applications, especially those that employs repository-based algorithms. This research tries to confirm the possibility of building language repository in Bahasa Indonesia involving user participation. We survey 191 people about their experience and opinion in using Google Translate, at the time they first used the application several years ago and recently. Opinions were then analyzed for sentiment score using lexicon-based sentiment analysis. The shift of sentiment was calculated and confirmed using t-test. There is a change in opinion score, from 0.3174 at first time use of the application to 0.4980 at recent use. The improvement can be attributed to the introduction of better translation algorithm and the improvement of data repository (or dictionary) by crowdsourcing. We conclude that it is possible to involve users to build language repository for Bahasa Indonesia.

1. Introduction
Repository is a place that collects or stores resources, and the term usually has implicit mean of immensity. The field of Natural Language Processing (NLP) has introduced more specific meaning for repository, which is usually a form of digital storage containing dictionary, corpus, or knowledge base. Corpus and knowledge base are collections of words, phrases or sentences that have been annotated with particular information [1]. Repository is important in NLP because many algorithms work upon its existence, for example algorithms that utilize WordNet [2, 3], N-gram[4], or lexicon [5, 6].

Repositories for Bahasa is unfortunately rare. Big dictionary may be the only repository in Bahasa. Thesaurus has been attempted to build, but the size is low. Other type of repositories such as WordNet, do not yet exist although attempts to build one have long been initiated [7]. Hence, implementation of repository based algorithms in Bahasa has not been successful.

Building a repository can be expensive because the work does not only involve digital data collection but also data annotation. Annotation is conducted by language experts and/or native speakers of the specified language. Repositories are traditionally built by a team of experts in a long-term project, which may involve people under supervision. Examples are the development of language dictionaries [8] and WordNet [9]. However, the advances of internet technology and virtual interaction among people have facilitated a way to build repositories involving people in the virtual world. This is commonly called crowdsourcing. Wikipedia has a successful story to build digital repository containing encyclopedic entries. Wikipedia has more than 3 million users that produce 45 million
articles in various languages [10]. The number of Wikipedia articles is growing with the rate of 20 thousands articles per month involving more than 100 thousand active users.

To overcome the problem of building language repository for Bahasa, we proposed a solution to involve people in an online system. This article describes the potential to build language repository for Bahasa Indonesia, taking the case of Google Translate (GT). At the current stage, this research tried to confirm whether crowdsourcing can be successful and improve the performance of an application that utilizes the repository. This study reveals that many Indonesian speakers believe that GT has been performing better in recent years. The achievement can be attributed to the introduction of crowdsourcing in building the repository.

2. Related Works
Efforts to build language repositories for Bahasa Indonesia have been conducted by many parties. The earliest attempt was conducted by Balai Bahasa which built Kamus Besar Bahasa Indonesia (Grand Dictionary of Bahasa). The first edition was printed in 1988 and was available online at 2008 [11]. The same institution has published thesaurus containing synonyms, hyponyms and meronyms [12].

Building word net for Bahasa has also been attempted by [7] replicating the model of English WordNet. It has 1203 synset and 1659 words. However, this process does not seem to be continuing and the repository is hardly accessible. The idea to build Indonesian corpora has been raised several years ago by [13] but the result has not been evident.

In the aforementioned works, repository building is conducted by team members of the corresponding projects. Attempt has been rare to involve public in building an Indonesian language repository. We can mention Oxford Dictionaries and Google Translate among websites that try to attract Indonesian people to contribute in building English-Indonesian dictionary. To the extent of our knowledge, no Indonesian institution has been conducting such an effort.

3. Method
To confirm whether building repository of Bahasa through user participation can be successful, we have selected Google Translate (GT) as a study. This is because GT has started to use crowdsourcing for a while, since 2014 and the resulted repository is used in another part of the application [14].

3.1. Data collection
We surveyed 191 people about their experiences dealing with the GT. Our respondents come from various background including students, civil servants, and private sector employees. Their ages start from 21 to more than 40.

| No | Questions |
|----|-----------|
| 1  | Have you ever used Google Translate? |
| 2  | When did you first access Google Translate? |
| 3  | What was the last time you use Google Translate? |
| 4  | What are your reasons in using Google Translate? |
| 5  | In the scale of 1 – 10, rate the accuracy of Google Translate in translating a word? |
| 6  | In the scale of 1 – 10, rate the accuracy of Google Translate in translating a phrase? |
| 7  | In the scale of 1 – 10, rate the accuracy of Google Translate in translating a short sentence? (a complete sentence having less than 5 words) |
| 8  | In the scale of 1 – 10, rate the accuracy of Google Translate in translating a complex sentence? |
| 9  | How many times in total do you use Google Translate? |
| 10 | Have you ever recommended your colleagues to try Google Translate? |
|    | Write your opinion about the translation result by Google Translate at the first time you tried its service. |
| 11 | Write your opinion about the translation result by Google Translate at the last time you accessed the service. |
We prepared a questionnaire containing 12 questions. The questions are grouped into three parts. The first part is to confirm that the respondent uses the software. The second part asks the respondent’s experience in using the software. The last part asks the respondent’s opinion about the software. (see Table 1). User’s opinion about the accuracy of GT translation is probed using 1 – 10 level scoring. On the other hand, user’s opinion about the improvement of GT translation after years of usage is probed using free-text questions.

3.2. Data analysis

The main purpose of the research is to indicate whether user involvement and participation in compilation and validation of language resource or repository can improve the usefulness of the application that uses the resource or repository. Usefulness in this context may be proved by the opinion of users about the accuracy of resource and the shift of their opinion about the output of the application.

The opinion shift may be detected by analyzing the sentiment of answers to question 11 and 12 of the survey questionnaire. Both questions ask respondents for opinion about the translation result of GT. However, the first question focuses on the translation result several years ago, while the latter focuses on the translation result recently. As depicted in Figure 1, sentiment analysis and t-test are used to detect the opinion shift.

![Data analysis diagram](image)

Figure 1. Data analysis

Sentiment analysis of user answers were conducted using a lexicon based method utilizing SentiWordNet [15]. Several steps have to be gone through. Our survey used Bahasa Indonesia and users answer questions in Bahasa, while the lexicon based sentiment analysis utilizes SentiWordNet, which is an English repository. Therefore, prior to conducting sentiment analysis, free-text user responses were translated into English. Thereafter, we analyze the sentiment of each response statement using SentiWordNet. The output of this series of processes is a sentiment score of each response statement.

Translated responses were preprocessed before performing sentiment scoring. There are two steps in preprocessing, i.e. tokenizing and POS Tagging. Tokenization of a response statement divides the statement into a token of words and symbols. Tokenization also helps the step of POS tagging. We use an existing library (Apache Lucene) to perform tokenization.

POS tagging is conducted after tokenization. It aims at getting the exact part of speech based on the context of the sentence. Part of speech for each word is very decisive in the searching process of sentiment score in SentiWordNet. We used Stanford POS Tagger to do POS Tagging process. At the end of this process, we’ll get a list of tokens and the part of speech for each token.

SentiWordNet has the sentiment score of each token in the list, either positive or negative sentiment. Sentiment score of a response statement is obtained by summing up the sentiment score of tokens of that statement. Tokens are treated as lexicon in SentiWordNet. Part of speech of words is used to choose relevant synset. For polysemous words, we choose a sense using First Sense method.

Sentiment scores of responses for question 11 and 12 become two sets of scores. The first set resembles the opinion of users several years ago at the first time they used GT and the second set resembles their opinion recently at the last time they used GT. The two sets can be presented into a table (see Table 2). The average score can be used to decide whether the opinion of users shifts more
positive or more negative. To confirm that there is an opinion shift, we use t-test to examine statistically whether the two sets are similar or not (in other words, the opinion from the two time point is the same or different). A change towards positive opinion suggests that the application is successful in doing what it should be doing.

Table 2. Output of sentiment analysis against respondent opinion

| Respondent | Opinion scores |  |
|------------|----------------|---|
|            | First time access | Recent access |
| respondent 1 | ...              | ...          |
| ...         | ...              | ...          |

| Average     |                |

4. Result and Discussion

4.1. Result

Most of our respondents are frequent users of Google Translate (GT), translating text from and to Bahasa Indonesia. Among 191 respondents, 164 people have used the web application frequently, i.e. more than 25 times. About 151 respondents, or 79%, started using the application more than three years ago and 39 respondents started just recently. Among the 151 respondents who started using GT three years ago, 138 (or 91%) are active users who still work with the application during the last three months. The statistics confirms that the respondents are reliable to give opinion about Google Translate in that most of them are active users for several years.

Our survey results in 191 pairs of responses where each pair contains statements about using GT at first time use and at recent use. Sentiment analysis generates sentiment score of each statement hence we’ve got 191 pairs of opinion scores, which are grouped into two sets of data. The first set is for opinion at first time use and the other set is for opinion at recent use. Figure 2 depicts the opinion scores for the two sets. Scores in the first set are depicted into red points and scores in the second set are depicted in into blue points.

Not all respondents state that GT is better recently than it used to be. As depicted in Figure 2, several respondents state that GT is not getting better (as indicated by blue point that is lower than red point). However, many respondents believe that the output of GT is better recently than it used to be several years ago (as indicated by blue point that is higher than red point). The average opinion scores of the two sets are 0.317 and 0.498, which means that in average the respondents have more positive sentiment during the use of GT recently than that when they first used the application.

Figure 2. Opinion score for the 191 respondents at first time use (red) and at recent use (blue)

The average opinion scores for the two sets of data suggest that there is a change in user opinion. To prove the significance of the change, however, we have to examine the data using t-test. The t-test checks whether two groups of data are statistically different. Paired t-test was used because each pair
of data was obtained from the same respondent and for that purpose we have used the data analysis feature obtained in Microsoft Excel.

Table 3 presents the result of paired t-test for the two sets. Variable 1 is the set of opinion score at first time use, which has a mean score at 0.3174, while variable 2 is the set of opinion score at recent use, which has a mean score at 0.498. The t-test results in t-value as high as 3.5865, which is higher than critical t-value, either for one-tail or two-tail test. The p-value is very low (less than 0.001) which suggests that the test result is not coincidental. The t-test result approves that the two sets of opinion scores are statistically and significantly different. Therefore, we can conclude that users have more positive sentiment on the recent output of GT as compared to their experience several years ago.

|            | Variable 1 | Variable 2 |
|------------|------------|------------|
| Mean       | 0.3174     | 0.4980     |
| Variance   | 0.2812     | 0.3318     |
| Observations | 191        | 191        |
| Degree of freedom | 190     |            |
| t Stat (t-value) | 3.5865 |            |
| P(T<=t) one-tail | 0.0002131 |            |
| t Critical one-tail | 1.6529 |            |
| P(T<=t) two-tail | 0.0004262 |            |
| t Critical two-tail | 1.9725 |            |

4.2. Discussion

Google Translate has been improving in recent years as admitted by respondents in our survey. This is particularly true for translation from and to Bahasa Indonesia. Many respondents are apparently happy with the application and this statement is supported by the fact that 133 of respondents, or 70%, have ever recommended colleagues or friends to use GT. Such improvement may come from the improvement of translation algorithm. Translation algorithm of GT reached a new milestone around late 2016 when Google introduced the use of neural machine translation. The method reduces the number of translation error by 60 percent as compared to the previous algorithm called phrase-based machine translation [16]. However, translation process relies heavily on a form of repository, called dictionary. Hence, the size and correctness of repository plays a significant role in the improvement of the translation result. No one can imagine machine translation without the help of dictionary.

We realize that sentiment analysis may provide flawed output because of three reasons, i.e. phenomenon of thwarted expectation, slang and non-standard language, and ambiguous words. The flaw may be exacerbated because our sentiment analysis involves language translation from Bahasa to English which may not result in exact translated words [15], 17, 18]. However, we may expect that the use of the lexicon-based sentiment analysis on a large set of statements increase the variance of sentiment scores but it does not significantly change the mean. Therefore, the t-test result can still be valid.

Natural Language processing for Bahasa Indonesia still lacks in development. Existence of repository is one of the reasons. Bahasa has already had a dictionary and thesaurus but the number of entries is still low. On the other hand, many NLP algorithms require repositories to work, e.g. for [19]. Even the development of NLP algorithm needs repositories as a standard test bed. Therefore, it is imperative for the field of NLP to build repositories, and a way to do it is crowdsourcing.
5. Conclusion
This study confirms that the potential to involve users in building repository, particularly for Bahasa Indonesia, is existent. Our case study suggests that Google Translate has improved and the improvement can be attributed to the introduction of better translation algorithm and better data repository (or dictionary) involving user participation. The improvement is supported by our respondents’ opinion. Respondents’ statement about the result of GT translation at first time use has an average sentiment score of 0.3174, while their statement about the translation at recent use has an average sentiment score of 0.4980. Hence, it is possible to involve users to build repository of Bahasa Indonesia.

References
[1] J. Pustejovsky and A. Stubbs, Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. “ O'Reilly Media, Inc.,” 2012.
[2] J. J. Castillo and M. E. Cardenas, “Using sentence semantic similarity based on WordNet in recognizing textual entailment,” in Advances in Artificial Intelligence--IBERAMIA 2010, Springer, 2010, pp. 366–375.
[3] H. Liu and P. Wang, “Assessing sentence similarity using wordnet based word similarity,” J. Softw., vol. 8, no. 6, pp. 1451–1458, Jun. 2013.
[4] I. Islam, E. Milios, and V. Keselj, “Text Similarity Using Google Tri-Grams,” in 25th Canadian Conference on Advances in Artificial Intelligence, 2012, pp. 312–317.
[5] T. Wei, Y. Lu, H. Chang, Q. Zhou, and X. Bao, “A semantic approach for text clustering using WordNet and lexical chains,” Expert Syst. Appl., vol. 42, no. 4, pp. 2264–2275, Mar. 2015.
[6] H. Thamrin and A. Sabardila, “Utilizing Lexical Relationship in Term-Based Similarity Measure Improves Indonesian Short Text Classification,” ARPN J. Eng. Appl. Sci., vol. 11, no. 22, pp. 13141–13145, 2016.
[7] D. D. Putra, A. Arfan, and R. Manurung, “Building an Indonesian wordnet,” in Proceedings of the 2nd International MALINDO Workshop, 2008.
[8] J. D. Anderson, “The Development of the English-French, French-English Bilingual Dictionary: a Study in Comparative Lexicography.,” 1971.
[9] P. Smrž, “Quality control for wordnet development,” in Proceedings of the second international WordNet conference, 2004, pp. 206–212.
[10] “Wikipedia: Size of Wikipedia.” [Online]. Available: https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia. [Accessed: 24-Aug-2017].
[11] D. Sugono, “Selamat datang di KBBI Daring.” 2008. [Online]. Available: http://badanbahasa.kemdikbud.go.id/kbbi/. [Accessed: 01-May-2016].
[12] Pusat Bahasa, Tesaaurus Alfabetis. Bandung: Mizan, 2009.
[13] R. Manurung, B. Distiawan, and D. D. Putra, “Developing an Online Indonesian Corpora Repository.,” in PACLIC, 2010, pp. 243–249.
[14] F. Lardinois, “Google Wants To Improve Its Translations Through Crowdsourcing.” 2014. [Online]. Available: https://techcrunch.com/2014/07/25/google-wants-to-improve-its-translations-through-crowdsourcing. [Accessed: 24-Aug-2017].
[15] E. W. Pamungkas and D. G. P. Putri, “An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia,” in 2016 6th International Annual Engineering Seminar (InAES), 2016, pp. 28–31.
[16] Y. Wu et al., “Google’s neural machine translation system: Bridging the gap between human and machine translation,” arXiv Prepr. arXiv1609.08144, 2016.
[17] J. Holmes, An introduction to sociolinguistics. Routledge, 2013.
[18] B. Ohana and B. Tierney, “Sentiment classification of reviews using SentiWordNet,” in 9th. it & t conference, 2009, p. 13.
[19] H. Thamrin and A. Sabardila, “Using Dictionary in a Knowledge Based Algorithm for Clustering Short Texts in Bahasa Indonesia,” in 2014 International Conference on Data and Software Engineering (ICODSE), 2014, pp. 1–4.