TELKOMNIKA, Vol.17, No.5, October 2019, pp.2321~2326
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v17i5.12005

Crowdsourcing in developing repository of phrase definition in Bahasa Indonesia

Husni Thamrin*[1], Gunawan Ariyanto[2], Irma Yuliana[3], Wawan Joko Pranoto[4]
[1,2,3]Universitas Muhammadiyah Surakarta,
A Yani St., Pabelan, Kartasura, Sukoharjo, Jawa Tengah, Indonesia, tel: (+62271)717417
[4]Universitas Muhammadiyah Kalimantan Timur
15 Juanda St., Samarinda, Kalimantan Timur, Indonesia, tel: (+62541)748511
*Corresponding author, e-mail: husni.thamrin@ums.ac.id

Abstract

Language repository is valuable as a reference in using the language, its preservation, and in developing and implementation of natural language processing algorithms. Bahasa Indonesia is one of natural languages that hardly has repository despite its large number of speakers and previous attempts to build ones. We devised a way to develop repository of phrase definition in Bahasa using a kind of crowdsourcing and investigated its implementation. An application add-on was inserted to an information system that manages final year projects of undergraduate students. The add-on invites students to participate in writing keyword definition and validating definition. Investigation in a period of six months reveals that about 25% of application users take parts into the voluntary activities either as definition writers and/or validators. During the period, about 1200 phrase definitions were added into the repository and in average each definition is validated by two participants. The activity is supported by users that are well aware of the tasks, and have positive perception about the work, despite different reasons that motivate their contribution.

***Keywords**: Bahasa Indonesia, crowdsourcing, repository*

1. Introduction

A repository is a place that stores resources containing an large size of data, usually with tools to access the data [1]. A language repository is a repository that stores language resources in various types, such as words, phrases, sentences, and paragraphs, which is stored and processed by electronic means. The language repository (or commonly called corpus) may be in the form of dictionary, thesaurus, or collection of annotated texts.

Language repositories are important as a reference in using the language and they can be valuable to preserve a language along with the cultural context [2]. In the field of natural language processing, language corpora plays important role to run applications that use language dependent algorithms. Corpora in the form of N-gram, for example, can be used in word and sentence similarity calculation [3, 4]. Language corpora are also essential as testing materials in the development of methods, such as those implemented in classification, speech recognition, and machine translation [5–7]. Supervised algorithms use corpora as training data, while unsupervised methods may use corpora as testing data to assess algorithm performance. Even language independent methods may need corpora in proving the applicability of the methods in a target language.

Bahasa Indonesia is the formal language in Indonesia so it has a large number of users as more than 200 million people reside in the country. Unfortunately, language repository in Bahasa is scarce. The largest open repository is in the form of Big Dictionary that is widely available in print and is also accessible online [8]. However, many definitions in Big Dictionary are outdated. A book containing thesaurus in Bahasa has been published but the number of items is small [9]. Many researches claim that they use data repository they have built during research analysis but most repositories are not openly available [10, 11]. Developing language repositories can be costly for data collection, annotation, and validation [12, 13]. Many parties have attempted to build repositories of Bahasa but the result seems to have been unexciting [14–17]. Experts or natural speakers may need to get involved to validate or annotate

Received December 9, 2018; Revised April 15, 2019; Accepted May 9, 2019

the items. This writing describes an effort to build repository of phrase definition in Bahasa Indonesia. The effort is conducted utilizing technology mediated social participation through "SiS", an application that manages undergraduate student in their final project. Such a model of repository construction is commonly called crowdsourcing, which describes an activity in taking a task once performed by employees and outsourcing it to a large network of people [18].

2. Research Method

2.1. Application

An application named SiS has been used to collect data from the crowd. The main purpose of the application is to manage undergraduate student final project. The long term use of the application make it potentially suitable for implementation of the crowdsourcing model. At the end of undergraduate study, a student will do a brief research project and write a scientific paper that explains the result. The project proposal shall consist of a title, paragraphs of description, and several keywords. During research activities, students have to write their progress in an online log book, for at least eight times.

A companion application add-on has been inserted into the main application. Every after the third time a student write a log, the add-on displays a small pop-up box to attract student attention. The pop-up box contains a message that invites students to participate in developing or validating repository items. The participation is not obligatory, which may eliminate uninterested parties to join and take part. When a student clicks a link in the pop-up box, the system will redirect to web pages that enable the user to contribute, either by writing a new definition of phrases or validating phrase definitions. Now because student scientific papers are written in Bahasa Indonesia, the resulting repository will be in Bahasa.

2.2. Data Collection and Analysis

The application is relatively new for the university, and the add-on is even later. After one semester running, we investigated the number of students served by the system and the number of students that contributes in crowdsourcing activities. More importantly, we calculated the number of phrase definitions, and the number of definition validation contributed by the crowd. Later, we investigated students with most contributions and invited them into a survey. Of 34 students invited, 16 students showed up and took part in the survey. We had prepared a questionnaire containing 22 questions asking about the application feature and student perception which need student conformations as shown in Table 1. The questions are grouped under 4 categories and were intended to observe the user knowledge and perception, product usability and user interest.

Table 1. Questionnaire to contributing users

Statement	
Knowledge	Perception
1. I am aware of the feature to contribute on keyword definition and definition validation, in application SiS.	1. The keyword contribution feature for keyword definition and validation has an important role in the application SiS.
2. I understand the purpose of the features related to keywords in SiS	2. The keyword contribution feature in SiS helps users to work on their thesis.
3. The keyword in a scientific article and thesis is no more than a less useful complement.	3. The keyword definition feature in SiS gives users the freedom to define according to their understanding (with their own sentences).
4. A scientific term or keyword needs to be explained in terms of the definition of terms.	4. Features of keyword contributions in SiS are relevant to the needs of the thesis completion.
5. A scientific term or keyword can only have one definition.	5. The keyword validation feature in SiS provides no option to users to give their opinions.
Usability	6. The keyword contribution feature in SiS makes it easy for users to express their opinions.
1. I believe that the keyword contribution feature in application SiS has benefits.	7. The keyword contribution feature gives me a non-material / satisfactory reward for my participation.
2. I understand the usefulness of keywords in a scientific article or thesis.	8. SiS support the achievement of good quality thesis.
3. The definition or meaning of a keyword is needed to learn the topic and scope of discussion in a scientific article and thesis.	Interest
4. Validation of keyword definitions is not needed.	1. I participate in defining / validating keywords because it affects score of my project.
5. Younger students will benefit from my contribution in defining or validating the definition of keywords.	2. I love getting rewards every time I participate in making definitions / validating keywords.
	3. I participated in defining / validating because I was asked by my supervisor.

The first category is about knowledge that drives the user into that feature. The second category is about user perception during the interaction with the system, including the keyword significance for their work (importance and relevance). The third category observes product usability whether user contribution gives benefits. The last category is user interest. The main purpose of the research is to indicate whether the information system may be employed to do crowdsourcing. Respondents need to state whether they strongly agree, agree, disagree, or strongly disagree on each statement in the questionnaire. The response were then metered using Likert Scale and each response is converted into one of the values 4, 3, 2, 1, correspondingly.

3. Result and Discussion

3.1. Result

After a year of implementation, the number of application users have grown and reached a steady number of about 1000 people. Users come and go because of the nature of student final project activity. Students do a final project for one semester in average and though some extends for another semester, they will eventually finish their work, pass the course, graduate and finally cease accessing the application.

Average users can be assumed to have access to the system for six months (or one semester), hence our analysis may be cropped into a frame of six months. We have selected a time frame from February until August 2018 for analysis. As the application was relatively new, the number of users were still growing from 205 in February to 1086 in August. The growth was quick in early phase of the semester and was very slow by the end of the semester. On the same period, the number users that participated as definers and validators were also growing, but with steadier and slower rate.

Not all users participate in crowdsourcing activities. In February, about 17% users contributed in writing definitions, i.e. being definers, and the proportion increased slowly and reached a value of 25% in August. For August, the percentage is equivalent to 271 users. Smaller proportion is seen for users that involved themselves in validation, which ranges from 12% to 17% of the total users as shown in see Figure 1.

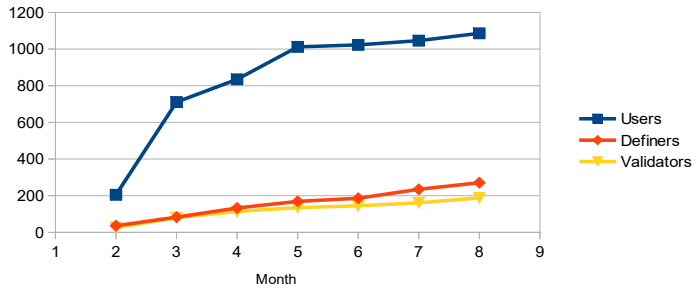

Figure 1. Number of users, definers and validators during the period of observation

Despite the smaller number of users to do the task, validation has more productive output than definition writing. The phenomenon has a straightforward explanation because validating a definition needs much less effort than writing a definition. A user just need to click a multiple choice option and click a button to select whether or not a phrase definition is accurate/acceptable or not. Figure 2 tells that the number of phrases increased by more than 3000 items during the period of February to August. The same period saw the increase of definition by about 1300 items while the number of validation is increased by more than 2700 items. The figures imply that about 40% of new phrases get definition from users and in average each definition is validated by two users.

Most of 271 users that contributed in writing definition did the activity for phrases in their own scientific paper, i.e. for their own project. Interestingly, there were about 34 people that wrote at least 7 phrase definition which means that they have written definitions from other

students' project. The latters were really engaged with crowdsourcing activity which is the reason for them to be invited to fill in a questionnaire for further investigation as why they were keen to contribute.

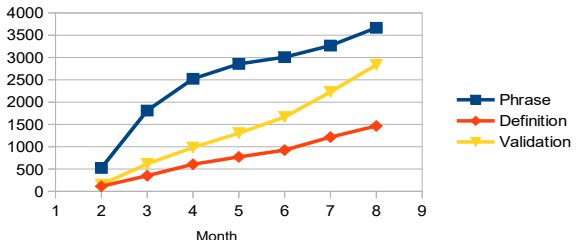

Figure 2. Number of phrase, phrase definition and validation during the period of observation.

Response of users against the questionnaire is displayed in Figure 3. Response for all category is more than 3 in Likert Scale which means that respondents are in average agree to strongly agree with statements in the questionnaire. Most users that were involved in crowdsourcing were well aware of the tasks. They knew what they were doing and what it was all about. They also have positive perception to the feature of defining phrases and validating definitions and feel somehow the feature is useful.

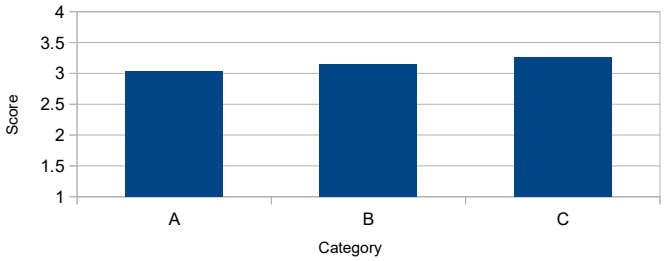

Figure 3. Response of users that were active in crowdsourcing;
Category A is knowledge, B is perception, and C is usability

Category D of the questionnaire asks about user intention in taking part in crowdsourcing activities. Among 16 students that filled in the questionnaire, ten students thought that they would get additional score in their final project by getting involved in the activities, seven said that they did it because their supervisor suggested them to. Most of them (or 13) said that they got a sort of reward by doing the good thing. We observed at least three students which did all the works purely for the sake of goodness.

3.2. Discussion

An information system that manages undergraduate student final projects has been deployed as well as the add-on to crowd source a repository of phrase definition. The student users make definitions of keywords that they use in their scientific paper as part of their final project. Aside from phrase definition, students are opted to validate definitions written by their friends. The activities of writing phrase definitions and validating definitions are not obligatory. Instead, a pop-up box shows up to users randomly and intermittently, inviting them to join crowdsourcing activities [19]. The pop-up box is like an open call to all registered users. The application gives a sort of reward in the form of virtual score and medals. Moreover, the application has a large number of active users, hence it fits the conditions to use in crowdsourcing.

The application has attracted up to 25% registered users to contribute. This number is well above an expected value of 10% users as predicted by [20] who stated that 90% of users are expected to be passive. We spotted approximately 3% of users give much higher contribution than average users. One user shows exceptional work by contributing 51 word definitions alone. The contribution figure is not comparable off course to the work by the Madman in Winchester's tale [21], but the result for us is not less important.

The number of repository items has been growing since the application add-on was deployed. Phrase definitions increases by about 200 new entries each month, though the growth is unsteady in the range of 150–250 items. On the other hand, definition validation increases progressively during the period of observation, but the increase appears to reach a steady value of 400 validation in further months (September and October). The result is stimulating. If the growth is steady at the aforementioned rate, we would optimistically have a dictionary of scientific terms with ten thousand entries after four years.

The optimism has a good reason based on the observed statistics. Success in building repositories in main international languages such as in [22–25] may be copied or conducted better. However, we need to get alert by the phenomena which is hard to estimate. Students who participate in crowdsourcing do not have single motivation. A few have stated that they got involved to follow directions of their supervisors. Some assumed that participation in crowdsourcing would add to the score of their final project. Some others have joined for immaterial rewards, i.e. a worthy cause. Apparently, some reasons may be unsustainable, which may disrupt target achievement.

Strategies may need to be thought up and implemented to keep the good work moving on. Experts have put some guidance for a variety of a crowdsourcing project. It should have a clear goal, a sound challenge, and regular report. The application should be easy and fun, reliable and quick, intuitive, and provide options to the user so they can choose what they work on [26]. Besides, the contributors should be acknowledged, rewarded, and trusted. The content should be interesting, novel, focused on history or science, and there should be lots of it to create through the years.

4. Conclusion

Language repository can be developed through a crowdsourcing application. We have developed such a system that fit the conditions for crowdsourcing activities. It has a large number of trustable users, i.e. approximately one thousand students per semester. The main system runs well and it provides facilities for users to have their final project done. About 25% users participate in crowdsourcing activities to make phrase definitions and to validate definitions. During the period of observation, about 200 phrase definitions were written and in average each definition was validated by two users. In short, it is possible to develop language repository, in this case: phrase definition in Bahasa Indonesia, using an application that implements crowdsourcing model.

References

[1] Thompson MPA, Walsham G. Placing knowledge management in context. *J Manag Stud*. 2004; 41: 725–747.
[2] Dimitrova L, Garabik R. *Bilingual Corpus-Digital Repository for Preservation of Language Heritage*. Proc. Int. Conf. Digit. Present. Preserv. Cult. Sci. Herit. DiPP. 2012: 32–41.
[3] Islam A, Inkpen D. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans Knowl Discov from Data*. 2008; 2(2): 10.
[4] Islam A, Inkpen D. Real-word spelling correction using Google Web IT 3-grams. Proc. 2009 Conf. Empir. Methods Nat. Lang. Process. 2009; 3: 1241–1249.
[5] Scannell KP. *The Crúbadán Project: Corpus building for under-resourced languages*. Build. Explor. Web Corpora Proc. 3rd Web as Corpus Work. 2007; 4: 5–15.
[6] Klimt B, Yang Y. T*he enron corpus: A new dataset for email classification research*. Eur. Conf. Mach. Learn. 2004: 217–26.
[7] Thamrin H, Pamungkas EW. A Rule Based SWOT Analysis Application: A Case Study for Indonesian Higher Education Institution. *Procedia Comput Sci.* 2017; 116: 144–50.
[8] Sugono D. Selamat datang di KBBI Daring 2008. http://badanbahasa.kemdikbud.go.id/kbbi/ (accessed May 1, 2016).
[9] Pusat Bahasa. Tesaurus Alfabetis. Bandung: Mizan; 2009.

[10] Asian J, Williams HE, Tahaghoghi SMM. *A Testbed for Indonesian Text Retrieval*. Proc. 9th Australas. Doc. Comput. Symp. 2004: 2–5.
[11] Pamungkas EW, Putri DGP. *An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia.* 2016 6th Int. Annu. Eng. Semin. 2016: 28–31.
[12] Fort K, Adda G, Cohen KB. Amazon mechanical turk: Gold mine or coal mine?. *Comput Linguist.* 2011; 37: 413–20.
[13] Hessel H. Tools of the Trade MTurk 101 : An Introduction to Amazon Mechanical Turk for Extension Professionals Uses of MTurk by Extension Professionals. *Journal of Extension.* 2016; 54: 0–7.
[14] Putra DD, Arfan A, Manurung R. Building an Indonesian wordnet. Proc. 2nd Int. MALINDO Work. 2008.
[15] Wicaksono AF. Automatically Building a Corpus for Sentiment Analysis on Indonesian Tweets. 2014.
[16] Manurung R, Distiawan B, Putra DD. *Developing an Online Indonesian Corpora Repository*. PACLIC, 2010: 243–249.
[17] Rahutomo F, Rohadi E. Development of the Indonesian Language Information Retrieval System Research Tool (in Indonesia: Pengembangan Piranti Penelitian Sistem Temu Kembali Informasi Bahasa Indonesia). SESINDO 2015. 2015.
[18] Marcus A, Parameswaran A. Crowdsourced Data Management: Industry and Academic Perspectives. *Found Trends ® Databases.* 2015; 6: 1–161.
[19] Thamrin H, Ariyanto G, Yuliana I, Purworini D. *An Application that Invites Users to Participate in Developing Repository of Bahasa Indonesia*. 2018 Int. Conf. Comput. Control. Informatics its Appl., LIPI. 2018: 72–76.
[20] Gatautis R, Vitkauskaite E. Crowdsourcing application in marketing activities. *Procedia-Social Behav Sci.* 2014; 110: 1243–1250.
[21] Winchester S. The Professor and the Madman: A Tale of Murder, Insanity, and the Making of the Oxford English Dictionary (hardcover). 1998.
[22] Kiyota Y, Nirei Y, Shinoda K, Kurihara S, Suwa H. *Mining User Experience through Crowdsourcing: A Property Search Behavior Corpus Derived from Microblogging Timelines*. Web Intell. Intell. Agent Technol. (WI-IAT), 2015 IEEE/WIC/ACM Int. Conf. 2015; 3: 17–21.
[23] Leemann A, Kolly MJ, Britain D. The English Dialects App: The creation of a crowdsourced dialect corpus. *Ampersand*. 2018; 5: 1–17.
[24] Asai A, Evensen S, Golshan B, Halevy A, Li V, Lopatenko A, et al. HappyDB: A Corpus of 100,000 Crowdsourced Happy Moments. 2018.
[25] Lardinois F. Google Wants To Improve Its Translations Through Crowdsourcing 2014. https://techcrunch.com/2014/07/25/google-wants-to-improve-its-translations-through-crowdsourcing (accessed August 24, 2017).
[26] Saxton GD, Oh O, Kishore R. Rules of Crowdsourcing: Models,Issues, and Systems of Control. *Inf Syst Manag*. 2013; 30: 2–20.