Technical Report

The First Research Dive on
Natural Language Processing
for Sustainable Development

October 2016

Australian Aid

PULSE LAB JAKARTA

Advisor Note

Supporting Research Development for Academia

This is “real diving”! For two days, researchers gathered to find research questions and their solutions. Their interests on similar research areas were able to unite them to work together. There are several interesting ideas and results that are useful and can be extended for further-research. Even though it’s sometimes tiring, everyone enjoyed the event. ot only ideas and solutions, there were also new friends, connections, and shared knowledge. All of these happened in these two days. In my opinion, this event is really suitable for Indonesia academicians since most of us have difficulties allocating special time focused on research due to the abundance of non-research activities at the university. I really hope that there will be more similar events to support research developments in Indonesia.

Dr. Eng. Ayu Purwarianti, ST., MT.
Advisor for Natural Language Processing

Dr. Eng. Ayu Purwarianti, ST., MT. is one of the two lecturers invited to be the advisor of *Research Dive Natural Language Processing and Linguistics*. As a lecturer from Institut Teknologi Bandung (ITB), Dr. Ayuis an expert on Natural Language Processing, Speech Processing, Intelligent Tutoring System, and Knowledge Management System. She obtained her doctoral degree in Informatics from Toyohashi University of Technology in Japan. Before that,she attended ITB for both her undergraduate and master’s degree on Informatics engineering. One of her most notable project experience isdeveloping an algorithm which uses twitter data to signal the occurrence of strikes and demonstrations.

Combining Data and Linguistics

As the only participant with a linguistics background, I feel very fortunate to have been involved in this Research Dive event. Even though I only played a minor role, as an adviser in language matters, this event allowed me to realize the pressing number of language problems in relation to the manufacturing of translation programs that needs to be prioritized and tackled as soon as possible.

The selection of data in the Research Dive event which is comprised of language corpus from social media is large enough to give an accurate representation of the real use of language in the community. Thus, it can be used as a good starting point in making translation programs. Nevertheless, we should remember that this data still does not fully represent the use of language so the translation program created based on this data needs to be tested with other data of different characteristics. The problem of translation quality is one that requires plenty of attention because it is a problem that is quite complicated.

Although this was the first time such an event was conducted and only lasted for two days, I think the outcome is pretty good, and even exceeded initial expectations. Well-designed activities, (for example dividing the participants into several groups with different tasks), availability of adequate data, seriousness of the participants in work, and positive working conditions established throughout the event are all key factors that contribute to the success of the Research Dive.

Dr. Suhandano, M.A.
Advisor for Linguistics

Dr. Suhandano, M.A. represented Universitas Gajah Mada (UGM) as the other Research Dive advisor with a linguistics background. As an Associate Professor in the Faculty of Literature and Culture, Dr. Suhandano is one of the senior lecturers in UGM. He obtained his doctoral degree from UGM’sDepartment of Linguistics, his master’s degree from Australian National University, and his bachelor’s degree from also from UGM’s Faculty of Literature. While facilitating the Research Dive discussion, Dr. Suhandano had the opportunity to sharevaluable insights regarding the translation challenges brought forth by the difference in structure and vocabulary of English as compared to Indonesian languages.

Executive Summary

We live in a complex world where the rapid exchange of goods, information, and ideas has brought opportunities and prosperity to many, but also precipitated a heightened vulnerability to systemic risks. Governments, more than ever, need to listen to the feedback and insights of citizens.

But in a world of data overload, how can governments structure the feedback from citizens, discern meaningful insights and prioritise policy responses? This question is all the more pertinent in a country such as Indonesia where over 700 living languages are spoken.

To begin to address this need and support computational research initiatives, Pulse Lab Jakarta developed Translator Gator, a people-powered language game which translates words from English to any of six common Indonesian languages, namely Bahasa Indonesia, Sunda, Minang, Melayu, Jawa, and Bugis.

The game enabled the Lab to compile user-created dictionaries of words related to the Sustainable Development Goals. These dictionaries assist our partners in carrying out automated analyses of social media, to better understand which issues matter to people, such as what they are saying about education, health, climate change, and other key development challenges.

Gaming proved to be a powerful and efficient way to tap into the 'wisdom of the crowd'. In just a few months, Translator Gator gathered more than 109,000 user contributions from hundreds of players across Indonesia. After casting the net wide to gather this valuable body of data, Pulse Lab Jakarta recently hosted a group of linguistic experts to dive deep into this data.

The Research Dive took place at Pulse Lab Jakarta on 22-23 July 2016, and carried the theme Natural Language Processing for Sustainable Development. Over the span of this two-day event, 19 computational linguistic experts and advisors from 18 different universities and government research institutions were invited to collaboratively explore and analyse the data. It also served as an opportunity for the selected computational linguists to network and to share expertise. Split into four groups, participants were tasked with assessing the quality of the translations, visualizing the data to make better sense of it, and filling in important translation gaps in some of the dictionaries.

After completing the task during the Research Dive, the four groups submitted extended abstracts which are presented in this technical report. The first group expanded the Indonesian corpus by enlarging the translations using morpho-syntax and evaluating Levenshtein Distance. The second group suggested a technique to complete the untranslated words by using the Indonesian translation as the pivot language to create a better translation pair from the existing data. The third group explained the classification of the correct and incorrect translations within the Translator Gator dataset, by evaluating the feature lists – frequency, vote up, vote down, and lifetime – as well as the combination that affect the translation results. The fourth group developed and analysed a pyramid visualization called MIDVIS, which enables quick understanding of the Translator Gator dataset that associated with the 17 Sustainable Development Goals.

Pulse Lab Jakarta is grateful for the cooperation of Institut Teknologi Bandung, Universitas Gajah Mada, Airlangga University, Bina Nusantara University, Indonesia Institute of Sciences, Institute of Statistics, Islamic State University of Malang, Muhammadiyah Jember University, Satya Wacana Christian University, Sepuluh Nopember Institute of Technology, State Polytechnic of Jember, Telkom University, Trunojoyo University, Udayana University. Pulse Lab Jakarta is also grateful for the generous support of the Department of Foreign Affairs and Trade of the Government of Australia, which enabled this research collaboration and many of the Lab's other activities to advance data innovation in development practice and humanitarian action.

October 2016
Pulse Lab Jakarta

Research Dive Participants

Advisors

Dr. Eng. Ayu Purwarianti, ST., MT	Institut Teknologi Bandung
Dr. Suhandano, M.A.	Universitas Gajah Mada

Researchers

Group 1 – Analysis of the corpus in Bahasa Indonesia

Achmad F. Abka	Indonesian Insitute of Sciences
Badrus Zaman	Airlangga University
Firdaus Solihin	Trunojoyo University
I Putu Gede Hendra Suputra	Udayana University
Yulyani Arifin	Bina Nusantara University
Imaduddin Amin	Pulse Lab Jakarta

Group 2 – Analysis of parallel corpora

Arie Ardiyanti Suryani	Telkom University
Bagus Setya Rintyarna	Muhammadiyah Jember University
Banu Wirawan Yohanes	Satya Wacana Christian University
Isye Arieshanti	Sepuluh Nopember Institute of Technology
Sari Dewi Budiwati	Telkom University
Muhammad Subair	Pulse Lab Jakarta

Group 3 – Analysis of informal expressions or the quality of translation

A'la Syauqi	Islamice State University of Malang
Adi Heru Utomo	State Polytechnic of Jember
Fika Hastarita Rachman	Trunojoyo University
Lya Hulliyatus Suadaa	Institute of Statistics
Novita Hanafiah	Bina Nusantara University
Satyaning Paramita	Pulse Lab Jakarta

Group 4 – Visualization of Indonesian corpora

Hendra Bunyamin	Christian Maranatha University
Iwan Njoto Sandjaja	Petra Christian University
Ni Made Satvika Iswari	Multimedia Nusantara University
Retno Kusumaningrum	Diponegoro University
Muhammad Rheza	Pulse Lab Jakarta

Table of Contents

Translator Gator: Enabling Computational Research in the Public Sector

Yulistina Riyadi
yulistina.riyadi@un.or.id

Imaduddin Amin
imaduddin.amin@un.or.id

Abstract

Translator Gator informs computational research by crowdsourcing the translation of a set of linguistic terms, which can then be used to mine text-based datasets for insights. Translator Gator is a game which generates dictionaries for research by randomly allocating users with four different tasks, Translate, Synonymise, Evaluate, and Classify. Based on a recent translation campaign, we published two datasets which include (1) User translation and (2) User synonyms, for the purposes of the recent Research Dive.

1 Introduction

Gaining insights from citizens' feedback to inform public policy, whether the feedback is expressed in an active or a passive manner, for instance from citizens' complaints to governments through official channels or on social media, requires a set of keywords (formally, a taxonomy) by topic, government priorities for example, to process large unstructured datasets in a computationally efficient way.

Our work is motivated by three main challenges when one attempts to monitor perceptions of public policies and programmes in Indonesia from social media. We have found that the public sector in some developing countries has sub-optimal knowledge and capacity to collect such a taxonomy. Second, on social media, many linguistic variations, such as jargon and slang, make building a list of keywords more challenging as words, context and, by extension, meaning change across regions. Lastly, given the rich linguistic and cultural diversity in Indonesia, where over 700 living languages are recognised, difficulties are posed in that many languages and dialects are used in different provinces and islands. For instance, 'electricity blackout' can be written in various ways even in the national Indonesian language, such as mati lampu, lampu mati, mati listrik, and pemadaman. It is worth noting that such linguistic variation issue can be observed in many countries.

In this paper, we describe Translator Gator, a people-powered language game which creates a dictionary of translated `taxonomies', while providing a number of configurable functionalities, which are applicable to different kinds of research as well as (non-profit) social projects.

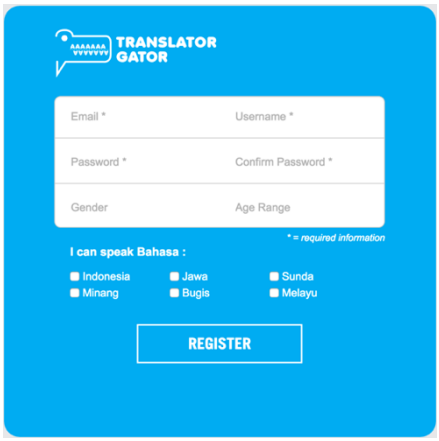

Figure 1. User registration page

1.1 Users

A new user engages in the game by clicking the register button, which leads the new user to a page where he/she is required to provide an email address, a user name, a password, and password confirmation. The user also needs to indicate the language(s) that s/he is fluent in, as shown in Figure 1.

2 Translator Gator

Translator Gator is designed and developed as a generic platform to crowdsource the translation and evaluation of a set of linguistic questions for supporting diverse computational linguistics research, not only offering functions for common tasks (such as translation, evaluation, and categorization), but also providing basic functions including rewards and quality control. It provides a mechanism to control the quality of results based on peer-review. To control the translation quality, users with bad reputations or, precisely, users whose translations are evaluated by others as `disagree' more than certain times, have their access restricted for a certain amount of time set by an admin. Thus, users need to be careful in providing answers. We plan to release Translator Gator as open-source software, with the aim of supporting computational research projects in the public sector and better monitoring programmes such as the Global Goals of the United Nations[1], by lowering the barriers to research concerning different languages through the development of more taxonomies.

[1] http://www.un.org/sustainabledevelopment/

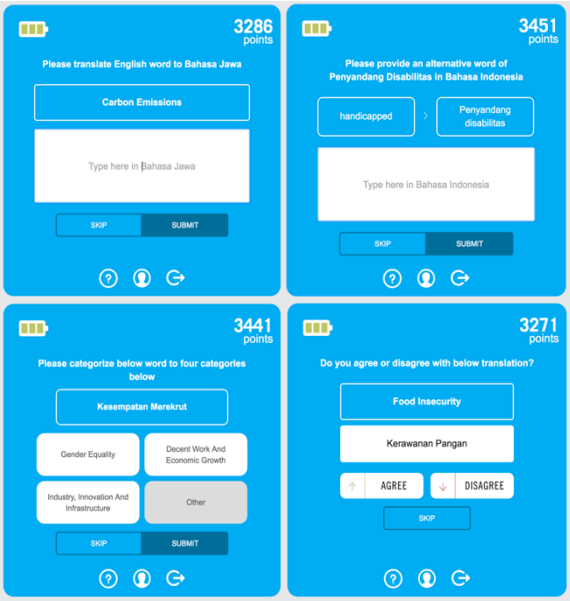

Figure 2. Translation task (top left), Synonymise task (top right), Classification task (bottom left), and Evaluation task (bottom right).

2.1 Four tasks

In this section, we explain the four main tasks Translate, Synonymise, Evaluate, and Classify. Translator Gator tasks a user in four ways, randomly based on a user's language fluency, which a user has to declare when subscribing to the system or can change its configuration in the user profile page. For instance, for a user who declares that s/he speaks Indonesian and Sundanese, the system asks him/her to translate English words to Indonesian and Sundanese, not other languages.

- Translate: Users are asked to translate words/phrases (e.g., access to water from the second Sustainable Development Goal) to everyday (dictionary and non-dictionary) words/phrases in other languages.
- Synonymise: Users are asked to suggest synonyms including non-dictionary expressions such as jargon, social media abbreviated expressions, and slang.
- Evaluate: Users are asked to evaluate the words/phrases contributed by others to validate the meaning as one of the following three, agree, disagree, and skip.
- Classify: Users are asked to classify words/phrases into predefined categories, allowing multiple choices. When the number of total categories is more than four, the system shows only three randomly chosen categories and 'Other'.

3 Data for the Research Dive

Under this initial Translator Gator project, we use an English taxonomy concerning the United Nations Post-2015 agenda, which was developed and compiled by United Nations Global Pulse[2].
We launched the crowdsourcing campaign between 25 January 2016 and 31 May 2016 and received 109,000 contributions across the four tasks. For the purposes of the Research Dive , we only examine data relating to three of the tasks which are translation, synonymise, and evaluation, because these datasets align with the field of Natural Language Processing.

We published two datasets including (1) user translation and (2) user synonyms for the purposes of the Research Dive. The user translation dataset consists of 38,762 user translation from 1,609 words in six languages and the user synonyms dataset consists of 18,403 synonyms for 7,801 translations.

3.1 User Translation Dataset

The user translation dataset describes the original words in English, the language of the translation, the user translations, the number of up votes and down votes of the translations, the user id and when the translation was submitted to the system, as described in the Table 1.

No	Entity	Value
1	word	Fewer jobs
2	language	Indonesian
3	translation	Pekerjaan lebih sedikit
4	vote_up	7
5	vote_down	1
6	anonymised user_id	5ab3456
7	timestamp	2016-01-25 16:08:12

Table 1. Sample of user translation dataset.

Table 2 shows the statistics of user translation transactions for each language. Figure 3 shows the achievement rate of the six languages in Indonesia, namely, Indonesian (IN), Jawa (JA), Minang (MI), Melayu (ME), Sunda (SU) and Bugis (BU).

Language	# users	# translated word	# translations	# vote-up	# vote-down
IN	331	1,609	33,873	19,946	2,898
JA	151	1,345	3,359	5,472	972
SU	57	463	595	1,471	669
MI	20	185	210	314	34
ME	29	415	542	821	82
BU	14	91	93	147	28

Table 2. User translation transaction of each language.

[2] http://www.unglobalpulse.org/projects/Post2015

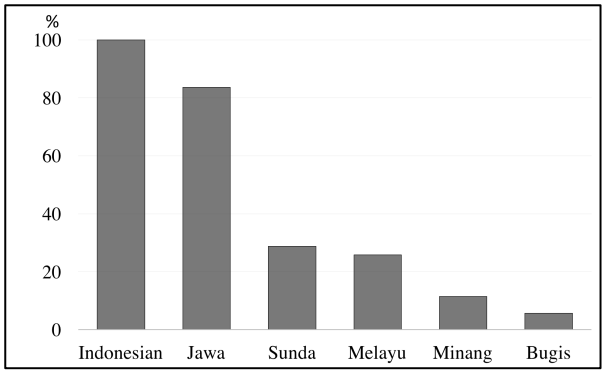

Figure 3. The achievement rate of the six languages, Indonesian, Jawa, Minang, Melayu, Sunda and Bugis.

3.2 User Synonyms Dataset

The user synonyms dataset describes the original words in English, the language of the translation, the original user translation, the synonyms, the user ID and when the translation was submitted to the system as described in the table below.

No	Entity	Value
1	word	pay cut
2	language	Indonesian
3	translation	Potongan pembayaran
4	synonyms	Pemotongan bayaran
5	user_id	5ab3456
6	timestamp	2016-01-25 16:11:22

Table 3. Sample of user synonyms dataset

Table 4 shows the statistics associated with user synonyms. Figure 4 shows the number of synonyms compared to the original translation.

Language	# users	# trans-lation	# synonyms	# unique synonyms
IN	313	7,556	18,017	10,547
JA	70	237	288	270
SU	7	37	41	39
MI	5	10	11	9
ME	10	36	46	45
BU	0	0	0	0

Table 4. The statistics of user synonyms

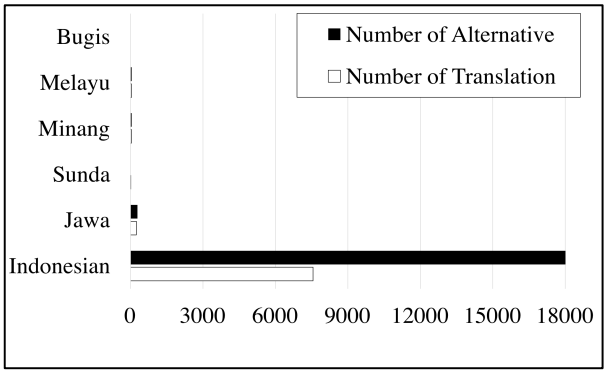

Figure 4. Number of synonyms per language

4 Summary

We built Translator Gator to begin to address the challenges associated with monitoring perceptions of public policies and programmes in Indonesia from social media. In Translator Gator, users are given four tasks, namely to translate, to provide synonyms/alternatives, to evaluate, and to classify development related words/phrases. In just a few months, Translator Gator gathered more than 109,000 user contributions from hundreds of players across Indonesia. For the purpose of Research Dive, we published two datasets including (1) user translation and (2) user synonyms. The user translation dataset consists of 38,762 user translation from 1,609 words in six languages and the user synonyms dataset consists of 18,403 synonyms for 7,801 translations.

Expanding Corpus by Enlarging Translation using Morphosyntax

Achmad F. Abka
Research Center for Informatics,
Indonesian Institute of Sciences
Bandung, Indonesia
achm056@lipi.go.id

Imaduddin Amin
Pulse Lab Jakarta
Jakarta, Indonesia
imaduddin.amin@un.or.id

Badrus Zaman
Universitas Airlangga
Surabaya, Indonesia
badruszaman@fst.unair.ac.id

Yulyani Arifin
Universitas Bina Nusantara
Jakarta, Indonesia
yarifin@binus.edu

I Putu Gede Hendra Suputra
Universitas Udayana
Badung, Indonesia
hendra.suputra@gmail.com

Firdaus Solihin
Universitas Trunojoyo
Madura, Indonesia
fsolihin@if.trunojoyo.ac.id

Abstract

This document describes our work on expanding corpus by enlarging translation using morpho-syntax. The corpus used in this work is from Translator Gator data. We use several methods to expand a list of synonymous words, such as by evaluating Levenshtein distance and structure of the text. The resulting synonym list will be compared to Google Translate.

1 Introduction

Translator Gator is a language game application that was developed by Pulse Lab to gather data keywords from Indonesian Language and other local Indonesian Language such as Bahasa Jawa, Sunda, Minang, Bugis, and even Indonesian slang. Translator Gator provides a database that contains a lot of keywords that can be used for academic and social research in Indonesia. The result from Translator Gator for the first 50 days can be seen in Fig.1.

Since results were gathered from volunteer participants, there needs to be a validation to make translation results more accurate. The system in Translator Gator provide vote up and vote down to recheck the result from the translation. The participants will vote up if they agree with the translation and they will vote down if they do not agree with it.

Participants of Translator Gator can give translations, vote, and suggest alternative translations. Because of that, the translation result can be a big corpus. There are some opportunities to enlarge the corpus using the similarity of the translated keyword and data from the feature vote in the Translator Gator.

The purpose from this research is to expand the translation that will be enlarge the corpus. The suggested method to be used is checking the similarity using the Levenshtein Distance and considering the syntax of Bahasa Indonesia D-M (diterangkan-menerangkan) (Ali Sjahbana, S. T, 1975).

Sanfilippo, A., & Steinberger, R. (1997) provide automatic selection and ranking candidates of translation using bilingual dictionaries that are enriched with thesaurus information. Castillo, J. J. (2010, August) uses the Machine Translation Systems to increase the size of Corpus to generate additional pairs translation using double translation process; for example translate one word in English into Spanish and then translate it back from Spanish to English. They use the double translation process to produce new pairs Text and Hypothesis. Sarkar, S.et al (2016) suggested three features to produce similarity, unigram over-lap count, normalized Levenshtein distance and the score result from Meteor Machine Translation.

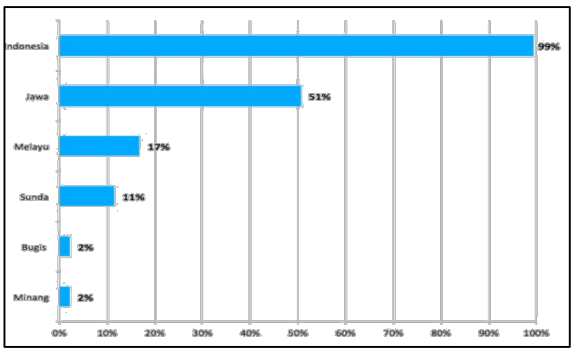

Figure 1. Translator Gator dictionary completion rates

2 Methods

Synonyms of translated English words, which is in Bahasa Indonesia, are collected. We collect the synonym from the same translation data of Translator Gator. The synonymous words are detected by looking at the text structure. If the same English word is translated to different words in Bahasa Indonesia, then we treat these Indonesian words as synonyms.

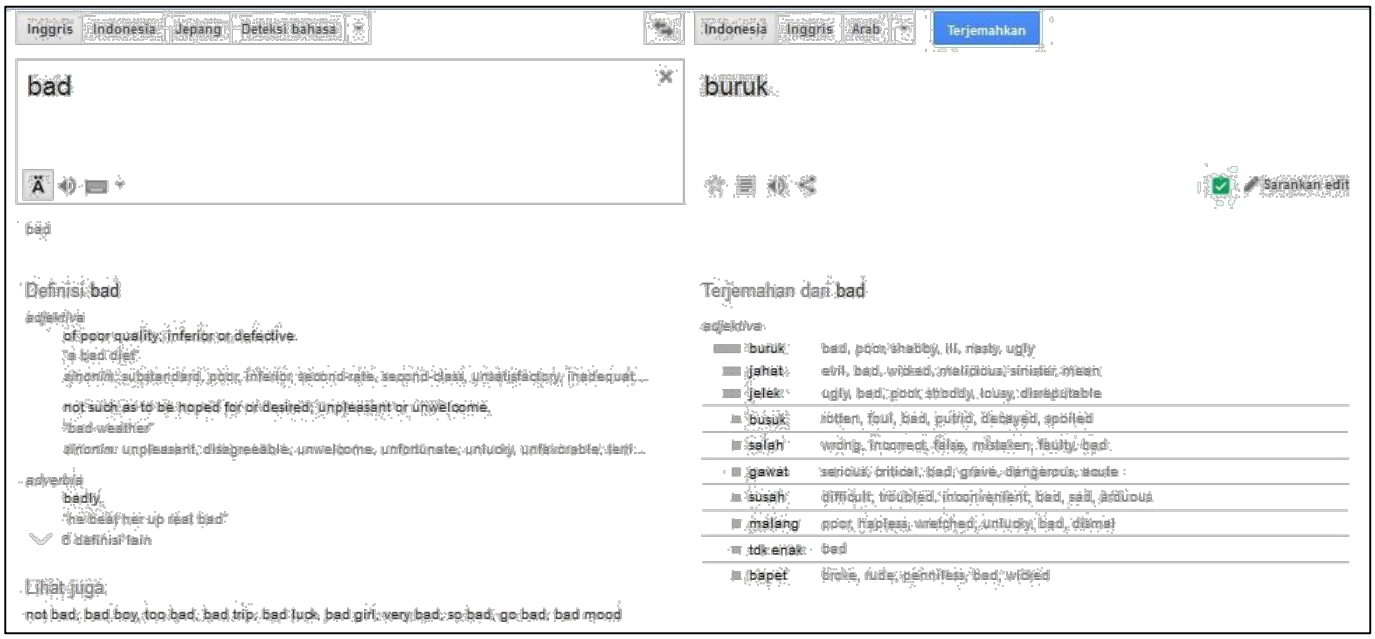

Figure 2. Google Translate result

For example, “equality” is translated to “keseimbangan”, “kesetaraan”, “persa-maan”, “kesamaan”, and “seimbang”. We treat these 5 words as synonyms.

The synonyms are also collected from phrases (limited to phrase with 2 words). The correct translation of each word is determined by its position. In Bahasa Indonesia there is a diterangkan-menerangkan (D-M) law (Ali Sjahbana, S. T, 1975). This law talk about the structure of the text. The described word is located in front of word that describes it. This is contradictory with English where the structure is menerangkan-diterangkan (M-D). By this law, we can examine a phrase to obtain translation or synonym of a word. For example, “food price” is translated to “harga makanan”. According to the D-M law, we know that the translation of “food” is “makanan” and the translation of “price” is “harga”. If “makanan” did not exist in the vocabulary, then we add “makanan” with its original word (“food”) to the vocabulary. If “food” already exist in the vocabulary with one or more translation, we add “makanan” as alternative of its translation (synonym of the current translation). The same treatment is also carried to the word “price” (“harga”).

Bahasa Indonesia also uses the concept of affixes. We also use levenshtein distance to recognize synonym of words. If two words have a levenshtein distance that is smaller than the defined threshold then they are treated as synonyms. For example, “bermain”, “main”, and “mainan” are treated as synonyms because they have a levenshtein distance of three or lower. Three can be chosen as threshold because a single affix in Bahasa Indonesia usually have length three or two. Although, multiple affix can be used in single word.

3 Evaluation

To evaluate our work, we want to compare the obtained synonym list with the result from Google Translate. Google Translate gives alternative translations of a word. It also gives information about the ranking of translations. It is somewhat shown by the scaled bar beside the words. This can be seen in Fig. 2. The word “bad” is translated to “buruk” and some alternative such as “jahat”, “jelek”, “busuk”, etc.

To compare the synonym of our work with Google Translate, we need to rank (give score) our synonym first. We want to score a word using the probability of word based on its surrounding words. We use the word class as the feature of surrounding words. The score of synonym word is its probability if the word class of the surrounding words are known. We use the same corpus from Translator Gator to count the probabilities. Only one word before or after the target word were taken into account in calculating the score (bigram model).

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
OD	Ordinal number
DT	Determiner / article
FW	Foreign word
IN	Preposition
JJ	Adjective
MD	Modal and auxiliary verb
NEG	Negation
NN	Noun
NNP	Proper noun
NND	Classifier, partitive, and measurement noun
PR	Demonstrative pronoun
PRP	Personal pronoun
RB	Adverb
RP	particle
SC	Subordinating conjunction
SYM	Symbol
UH	Interjection
VB	Verb
WH	Question
X	Unknown
Z	Punctuation

Table 1. Bahasa Indonesia Tag set

Before we can count the probabilities of synonym word, we need to tag every word with its

word class. To do that, we use Rashel et al. (2014) part-of-speech (POS) tagger. The tag set of Rashel tagger can be seen in Table I. Example of resulting tagged text can be seen in Table II. After each word has been tagged with its class tag, then we generate the bigram. To generate the bigram we only use translation that contain more than one word. Example of the bigram can be seen in Table III. The probability of a word is calculated using this bigram data. The ranking of each synonym is based on this probability. Word with higher probability assigned with higher rank.

No.	Tagged Text
1	3G/CD ./Z
2	tiga/CD batu/NN berapi/VB ./Z
3	Koneksi/NN internet/NN tanpa/SC kabel/NN sangat/RB mahal/JJ ./Z
4	sekolah/NN miskin/JJ ./Z
5	Memukul/VB perempuan-perempuan/NN ./Z

Table 2. Sample of text with tag for each word

No.	1st Word	2nd Word	1st Tag	2nd Tag
1	Koneksi	3G	NN	CD
2	Partisipasi	warga	NN	NN
3	terluka	ketika	VB	SC
4	Antar	negara	VB	NN
5	Kampus	abal	NN	X

Table 3. Sample of bigram generated from Translator Gator data

4 Conclusion

Enlarging Corpus with expanding the translation result from English to Indonesian can use the similarity checking. Using checking the similarity by Levenshtein Distance and also check the structure in Indonesian Language can give more reliable result translation. The meaning from each word in Indonesian sentence will be different depend on the position in the sentence or tagger.

For the further research the expand translation will be checking for more than two words or phrase.

References

Alisjahbana, S. T. (1975). Tatabahasa baru bahasa Indonesia (Vol. 1). Dian Rakyat.

Castillo, J. J. (2010, August). Using machine translation systems to expand a corpus in textual entailment. In International Conference on Natural Language Processing (pp. 97-102). Springer Berlin Heidelberg.

Rashel, F, et al. (2014). Building an Indonesian rule-based part-of-speech tagger. Asian Language Processing (IALP) 2014 International Conference on. IEEE.

Sarkar, S., Das, D., Pakray, P., & Gelbukh, A. (2016). JUNITMZ at SemEval-2016 Task 1: Identifying Semantic Similarity Using

Levenshtein Ratio. Proceedings of SemEval, 702-705.

Sanfilippo, A., & Steinberger, R. (1997). Automatic selection and ranking of translation candidates. In Seventh International Conference on Theoretical and Methodological Issues in Machine Translation: TMI (Vol. 97, pp. 200-207)

Enriching English into Sundanese and Javanese Translation List Using Pivot Language

Arie Ardiyanti Suryani
Telkom University
rie006@yahoo.com

Isye Arieshanti
Sepuluh Nopember Institute of Technology
isye.arieshanti@gmail.com

Banu Wirawan Yohanes
Universitas Kristen Satya Wacana Salatiga
banu.yohanes@staff.uksw.edu

Muhammad Subair
Pulse Lab Jakarta
muhammad.subair@un.or.id

Sari Dewi Budiwati
Telkom University
saridewi@tass.telkomuniversity.ac.id

Bagus Setya Rintyarna
Muhammadiyah Jember University
bagus.setya@unmuhjember.ac.id

Abstract

This paper discusses the problem of sparse translation of English into Sundanese and Javanese that were found in Translator Gator. Translator Gator is a language game created by Pulse Lab Jakarta, to support the research initiatives in Indonesia. Thousands of keyword were generated and translated from English into some Indonesian local languages using the crowd resource. Unfortunately, many English words are still has no translation in Javanese as well as Sundanese. To overcome this problem we propose a technique to fill the un-translated English words in Javanese and Sundanese using Indonesian translation as a pivot language. Evaluation was made by manually investigated whether each phrase results a proper translation. Experiment shows that our technique results relatively low translation accuracy. Limited coverage of phrase translation list and ambiguous words are identified as causes of translations errors in our technique.

Keywords—pivot language, translation weight, phrase translation.

1 Introduction

Parallel corpus is a collection of text in one language and their equivalent translation to other language. In machine translation research area, some language pairs contain a large number of parallel corpus are easy to obtain and ready to use. Conversely, for many languages pairs with a low resources language, there only a few of parallel corpus in small scale or even not found at all. The sparse of parallel corpus directly will result to a poor translation.

Similar problem faced by the Translator Gator, an online language game created by Pulse Lab Jakarta. It was built to collect a large number of keyword adopted by the 17 Sustainable Development Goals (SDGs). These keywords were firstly defined and translated into Indonesian by using the Google Translate. The crowd then translated these keywords into some Indonesian local language, such as Sundanese, Javanese, Buginese, and Minangnese. To attract many people to translate actively, the Translator Gator was represented as an online game, thus there was a reward and penalty. People will get some points when their translation was agreed by other (vote-up), otherwise they will lose their points (vote-down) as well as can be banned from continue playing at certain limit. A certain number of accumulative points were then can be redeem with a phone cell credit. In further, these translated keywords will be used to disseminate crucial information of food resilience, global warming, public health, as well as to be used by researches in computational linguistics and some related areas.

By the end, Translator Gator collected more than 160 million of keywords. These keywords either can be a single word or a phrase contains more than one word. All of the keywords translated into Indonesian completely. Unfortunately, only 80% and 20% of these keywords were translated into the Javanese and the Sundanese respectively (Riyadi & Amin, 2016). Therefore, we proposed a technique to enrich the translation list of English into Javanese as well as Sundanese using Indonesian as a pivot language. A pivot language was being chosen as a solution because English and Indonesian local language pair has limited resources, such as parallel corpus, dictionaries, and other language tools. We hypothesized that using the existing Translator Gator data is a reasonable solution so that it can be implemented immediately.

Our technique comprises of three sequential steps. Firstly, three pairs of translation terms are chosen. Those pairs are English-Indonesian translation, English-Javanese Translation, and English-Sundanese translation. Since one English term can be translated into several terms in Indonesian, Javanese and Sundanese, we choose

only one translated term according to the weight. The weight calculation involves vote-up, vote-down and frequency of translated result. Secondly, Indonesian-Javanese and Indonesian-Sundanese dictionaries are generated using Moses Translation System (Koehn, 2015). At last, a rule-based technique is employed to fill the un-translated English terms into Javanese and Sundanese terms. In the rule-based model, the process requires keyword label whether the word is a borrowed word or not. Thus for this step we employ a borrowed word list collected from the online resources. The translation result is then evaluated manually to observe how well our technique produces the translation.

On the next section we present the related work, short descriptions of Translator Gator, enrichment techniques, detail of the proposed technique, experiment results and finally enclosed with a conclusion.

2 Related Work

Previous related researches focus on the problem of text translation from one source language into a target language by using an intermediate (pivot) language. According to (Wu & Wang, 2009), there are three different pivot translation techniques that are triangulation method, transfer method and synthetic method.

The first method trains the source-pivot and pivot-target translation model by using parallel corpus. Using these two model, the translation model of source-target is then induced. The triangular technique was used by (Cohn & Lapata, 2007) to solve the small data size problem in English to French translation by using Dutch, Danish and Portuguese as an intermediate language. The Experiments show an improvement of translation results compare to the standard phrase-based translation. One of the problems raised in the triangular method is a very large resulted translation model and some phrase pair that might be not connected to each other because does not have the same pivot phrase (Cui, Zhu, Zhu, & Zhao, 2015).

The second method is transfer method that translates a source into target text in two consecutive steps that are source to pivot and pivot to target translation. A sentence of a source language is firstly translated into N pivot sentences, and then each pivot sentences translated into M target sentences (Utiyama & Isahara, 2007). The translation result is selected by using a defined weighting mechanism. Experiment shows that the transfer method was inferior to the triangular method for an English-Germany translation by using French as intermediate (Utiyama & Isahara, 2007).

The third method is synthetic method, which creates a new parallel corpus of source-target by translates pivot sentences into sources sentences using source-pivot translation as well as translated pivot sentences into target using pivot-target translation. This method applied by (Gispert & Mariño, 2006) for Catalan-English using Spanish as pivot. The evaluation of this paper was performed through comparing the translation result with or without the synthetic method. The experiment shows that the translation resulted were slightly inferior to the baseline. Other experiment was also employed by (Klementiev, Irvine, Callison-Burch, & Yarowsky, 2012) for English-Spanish translation. They use a large number of English and Spanish monolingual corpora and a small size of dictionary.

Generally, our proposed technique adopts the first idea. We attempt to create the pivot-target translation table by using the existing target-pivot translation list. The resulted translation table is then being used to translate source-target keywords that are still having empty translation.

3 Proposed Technique

There are three sequential step to fill the empty translation of Sundanese or Javanese, that are selecting English-Indonesian (EN-ID) phrase pairs, create Indonesian-Javanese (ID-JW) and Indonesian-Sundanese (ID-SU) dictionary as pivot, and then translating the empty Javanese and Sundanese translation Block Diagram of these steps shown in Figure 1.

The first step was intended to pick only good enough EN-ID translation pair, based on the number of translation result, the number of other user that agreeing this translation (vote-up), and the number of disagree user (vote-down). For this purpose, we define a weighting formula to pick EN-ID translation pair as shown in Equation 1.

$$\forall x : maximum(weight(y)) \tag{1}$$

$$weight = \sum y + \sum voteUp_y - \sum voteDown_y$$

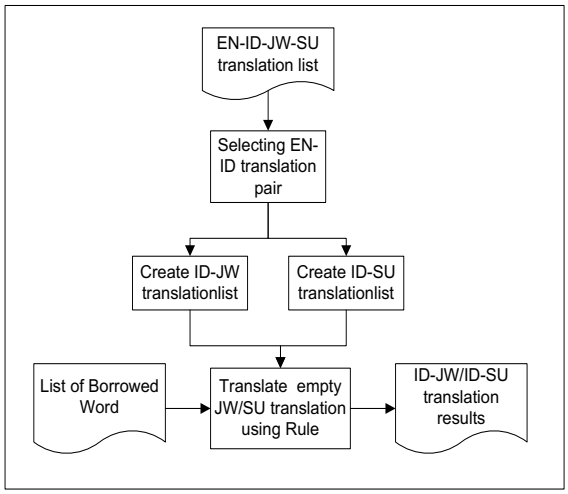

Figure 1. Our Proposed Technique Block Diagram

Given translation list consists of a number translation pair. For each translation pair x, we calculate the weight for each translation alternatives y. Selected translation pair was the one that has highest score among other translation pair. Whereas weight determined by frequency of each translation alternatives, then added by its number of vote-up and subtracted by its vote-down.

To create ID-JW and ID-SU translation list, previously we applied the same equation 1 to the EN-JW and EN-SU translation list. After that, ID-JW and ID-SU translation list was created by joined the EN-ID resulted by step 1 with ID-JW and ID-SU respectively. Unfortunately the translation list of ID-JW and ID-SU were only contains phrase, whereas translation of either single word or combination of word that build the phrase are not covered by ID-JW and ID-SU. Therefore, we do an enrichment of ID-JW and ID-SU dictionary by assumed them as parallel corpus and passed them into Moses translation system (Koehn, 2015), and took the translation model as our final ID-JW and ID-SU dictionary.

On the third step, the empty translation of Javanese and Sundanese are completed by using final ID-JW and ID-SU dictionary and applying a rule set as shown in Table 1.

Rule #1:
if (keyword is found in phrase translation table) **then**
  return the translation result
else apply rule#2

Rule #2 :
if (keyword is a borrowed word) **then**
  Javanese or Sundanese = Indonesian
else { keyword is not a borrowed word**}**
  if (keyword is a single word) **then**
    return "UNK" {UNK = unknown word}
  else {keyword is a phrase}
    split keyword into N words
    for each 1 until N word apply Rule 1-2

Table 1. Translation Rule

Translation result evaluation size was defined using Slovin formula [(Almeda, T. Capistrano, & G. Sarte, 2010). The sample size is define using the formula in equation 2, by using 5% error assumption for a given N total population.

$$sample\ size = \frac{N}{1+Ne^2} \qquad (2)$$

4 Experiment and Discussion

This part contains our experiment to observe the proposed technique performances, systematically started by description of experiment scenarios, dataset, experiment result and discussions.

A. Experiment Scenario and Dataset

There is only one experiment scenario that is to fill the Javanese or Sundanese translation for a number of keywords. The translation result was then evaluated manually to check their conformity. However, we also evaluate the initial ID-JW and ID-SU translation list resulted from applying equation 1, to ensure that this initial dictionary is quite good to be used as our dictionary.

In this experiment, we use 36.313 transactional data translation that translated by more than 100 Translator Gator users. It consists of 1324 unique English keyword. By using these keywords we get 1340 pair of initial ID-JW dictionary and 460 pair initial ID-SU dictionary.

B. Experiment Result & Discussions

Subjective evaluation to initial ID-JW and ID-SU dictionary were applied manually to all the ID-SU dictionary entry, and 40% of ID-JW. Result of this evaluation shown in Table 2.

Language Pair	Number of Evaluated Data	Translated Proplerly (%)
Indonesian – Sundanese (ID-SU)	460 of 460 (100%)	70%
Indonesian – Javanese (ID-JW)	540 of 1340 (40%)	68%

Table 2. Phrase Pair Evaluation

According to Table 2, the ID-SU and ID-JW were quiet good to be used as our dictionary to translate empty Sundanese or Javanese translation.

In the initial ID-JW and ID-SU dictionary, some error translation were found, one of them was caused by improper translation of a borrowed word, such as the word "*insiden obesitas*" (obesity incidence) and "*es dunia*" (global ice) that should be translated using Indonesian, rather than produces improper Javanese or Sundanese translation. The other error was caused by the translation that filled improperly by the user. In this paper, we overcome the first cause error by copying the translation if the keyword is a borrowed word. The resulted phrase pair was then used to translate empty Sundanese and Javanese translation.

The translation produced 269 Javanese phrase translation, consist of 76 translations generated using rule#1 and the other 193 translations resulted using rule#2. Whereas for Sundanese keywords, there were 1149 translations, comprises of 261 translations produced using rule#1, and 888 translations created using rule#2. It means that both ID-JW or ID-SU are majorly translated using rule#2. There are 77% of ID-SU and 71% of ID-JW were translated using rule #2, while only small portions of them are translated by using rule #1. It shows that our dictionary that was built using existing translator gator translation list covers less than 30% of the whole translation. Interestingly, we found that rule#1 gives better translation than rule#2 as depicts in Table 3. The translation results of the rule#1 achieve more than

fifty percent for both Indonesian-Javenese and Indonesian-Sundanese. Whereas using rule#2 both language pairs give translation result less than 35%. It means that although our phrase translation list covers only small portion of keywords it gives significant results in translation.

In this experiment, we evaluate all translation results of ID-JW. While for ID-SU translation we used only some translation sample determined using Slovin formula of ID-SU since a large number translation thus its relatively hard to evaluate all of these translation manually.

Rule Applied to translation	**Number of Evaluated Samples**	**Percentage of keywords translated properly**
Indonesian – Javanese (ID-JW)		
Rule #1 (using phrase translation / dictionary)	76	58 of 76 (76.3%)
Rule #2	193	42 of 193 (22.6%)
Total ID-JW evaluated sample	269	
Indonesian – Sundanese (ID-SU)		
Rule #1 (using phrase translation / dictionary)	68	60 of 68 (88%)
Rule #2	229	78 of 229 (34%)
Total ID-SU evaluated sample	297	

Table 3. Translation Result Evaluation

Furthermore, we also analyze some translation error produced in this translation. Since the rule #2 raises translation error majorly, our error analysis focus more on the translation error produced by rule#2. In addition, the translation errors generated in rule#1 commonly are caused by the low coverage of the phrase translation. Overall, translation error occurs in applying rule#2 was caused by two factors. The first factor is the incomplete or limited coverage of phrase translation list, whereas the second factor is the existence of ambiguous word.

We present some examples of translation errors arise in rule#2 in Table 4. The Source Phrase in the first column represents the Indonesian keywords that will be translated, while the target phrase refers to Javanese or Sundanese translation generated by our technique.

Actually these two problems might be minimized when bigger and more variety of parallel text added to the phrase table thus the possibility of a word occurs in the phrase translation list is higher.

Source Phrase	Target Phrase (Translation Results)	Occurs in	Error Type
menyogok	UNK UNK	ID-SU	words
pemilihnya (bought his votes)	(resulted no translation)		"*menyogok*" and "*pemilihnya*" does not exist in phrase translation list
energy ramah lingkungan (clean energy)	*Energy* UNK *lingkungan*	ID-JW	word "*ramah*" does not exist in phrase translation list
Sistem kesehatan yang layak (decent health systems)	*Sistem kasugengan ingkang layak*	ID-JW	not proper translation of the word "*kesehatan*" in phrase translation list
Sekolah yang buruk (bad school)	*Sakola ku awon*	ID-SU	not proper translation of the word "*ku*" in phrase translation list

Table 4. Some Error Examples Produced in Translation

Another solution of this problem is to use an Indonesian-Javanese and Indonesian-Sundanse dictionary to fill the unknown word resulted by this system. A morphological analyzer could also be added to smooth the translation result of an affixed word.

5 Conclusion

We proposed a technique to fill an empty translation from English into Javanese and Sundanese in the phrase translation list of Translator Gator System. We employed the existing phrase pair by considering Indonesian translation as a pivot to create English into Javanese or Sundanese translation.

Our experiment shows that the proposed technique by our team does not provide the accurate translation. We found that averagely our technique only reach 37% correct translation result of Indonesian-Javanese and 46% of Indonesian-Sundanese translation. However, this technique gives contribution to create a better translation pair from existing Translator Gator data, which gives more than 65% proper phrase translation for both Indonesian-Javanese and Indonesian-Sundanese pair translation. In the future, using dictionary to entail the translation quality is preferable.

Acknowledgement

We would like to thank to Pulse Lab Jakarta for introduce us to the problem discussed in this paper, as well as for provides data used in this experiment. A great appreciation also addressed to Telkom University for supporting of this paper publication.

References

Almeda, J., T. Capistrano, & G. Sarte. (2010). *Elementary Statistics*. Quezon City: UP Press.

Chahuneau, V., Schlinger, E., Smith, N. A., & Dyer, C. (2013). Translating into Morphologically Rich Languages with Synthetic Phrases. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1677–1687* (pp. 1677-1687). Seattle, Washington, USA: Association for Computational Linguistics.

Cohn, T., & Lapata, M. (2007). Machine Translation by Triangulation: Making Effective Use of Multi Parallel Corpora. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (pp. 384-355). Association for Computational Linguistics.

Cui, Y., Zhu, C., Zhu, X., & Zhao, T. (2015). *Augmenting Phrase Table by Employing Lexicons for Pivot-based SMT.* Arxiv.org.

Gispert, A. d., & Mariño, J. (2006). Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish. *In Processdings of LREC 5th Workshop on Strategies for developing Machine Translation for Minority Languages*, (pp. 65-68).

Klementiev, A., Irvine, A., Callison-Burch, C., & Yarowsky, D. (2012). Toward Statistical Machine Translation without Parallel Corpora. *EACL '12 Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 130-140). Avignon, France: Association for Computational Linguistics.

Koehn, P. (2015). *Moses-Statistical Machine Translation User Manual.* Edinburgh: University of Edinburgh.

Riyadi, Y., & Amin, I. (2016, June 30). *Translator Gator : Phase I Wrap Up.* Retrieved July 31, 2016, from United Nations Global Pulse: http://unglobalpulse.org/news/translator-gator-phase-I-wrap-up

Utiyama, M., & Isahara, H. (2007). A Comparison of Pivot Methods for Phrase-based Statistical Machine Translation. *Proceedings of NAACL HLT 2007* (pp. 484-491). Rochester, New York: Association for Computational Linguistics.

Wu, H., & Wang, H. (2009). Revisiting Pivot Language Approach for Machine Translation. *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP* (pp. 154-162). Singapore: Association for Computational Linguistics.

Discover Best Feature Combination of User Behavior in Indonesian Corpus Collection Based Incentive Crowdsourcing

Lya Hulliyyatus Suadaa
Institute of Statistics
lya@stis.ac.id

Novita Hanafiah
Bina Nusantara University
Nhanafiah@binus.edu

A'la Syauqi M. M. I.
Islamic University of Malang
syauqi@ti.uin-malang.ac.id

Adi Heru Utomo
State Polytechnic of Jember
adiheruutomo@polije.ac.id

Fika Hastarita R
University of Trunojoyo
fika@trunojoyo.ac.id

Ni Luh Putu Satyaning P.P.
Pulse Lab Jakarta
ni.paramita@un.or.id

Abstract

The research aims to classify translation correctness of translator gator dataset obtained from crowd-sourcing to evaluate features that affect translation. The nine features used for classification are Number of vote up and vote down, Frequency of word, Lifetime in seconds, Class (Correct or incorrect translation), Weighted vote up (number of vote up/lifetime), Weighted vote down (number of vote down/lifetime), and Diff score (weighted vote up-weighted vote down). Based on feature selection testing using Naïve Bayes, Random Forest, SVM, and J-48 method, it can be concluded the frequency feature can be combined with the weighted vote up and vote down for classifying the correct and incorrect translation. Diff score as differentiation of weighted vote-up and vote-down feature contributes poor result. However the combination of frequency and diff score using random forest method gives more accurate results with correctness percentage of 80.01% and can be used as alternate to classified the correct and incorrect translation.

1 Introduction

Translator Gator is a people-powered language game which creates a dictionary of translations to support academic research and social projects in Indonesia. It aims to translate a set of English keywords into Indonesian, as well as other local languages such as Bahasa Jawa, Sunda, Minang, Bugis, and even includes slang. The translated keywords will be applicable to many projects, especially those related to digital text analysis.

The problem in this research is how to classify the results of the translation obtained by the crowdsourcing. Many methods can be performed to evaluate the translation. One method that can be used is classification, one of data mining techniques.

The dataset used in this research is a dataset of translator gator. There are nine attributes / variables / features used to classify the translation, which will be grouped into two classes, each of which will be categorized in a group of words or user. These features are as follows: Number of vote up and vote down, Frequency of word, Lifetime in seconds, Class (Correct or incorrect translation), Weighted vote up (number of vote up/lifetime), Weighted vote down (number of vote down/lifetime), and Diff score (weighted vote up-weighted vote down)

In crowdsourcing data, there are some visible features and hidden features that affect classification method. Therefore, feature selection needs to be done. In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection is done by selecting the relevant features that affect the classification. Selection feature is used to reduce the dimensionality of data and features that are not relevant, as well as to improve the effectiveness and efficiency of the performance of classification algorithms. In this research, feature selection was done using J48, Random Forest, SVM, and Naive Bayes. This research aims to obtain a combination of features that can be used to classify the results of the translation obtained by Translator Gator.

2 Data and Methodology

Translator Gator application collects information related to the translation from source language (English) to target language chose by the user (Indonesia, Sunda, Jawa, Melayu, Minang, and Bugis). In addition, the game provides more features to validate the meaning of translation, such as evaluation function of translations submitted by others and suggestion of alternative words. These two features generates additional information that is used in the experiment, which are number of people who agree (vote up) and disagree (vote down) to the translation, as well as the lifetime of the words in seconds. The lifetime variable is gotten from how

long the translation data has been stayed in the database. The processing phase consists of three steps: preprocessing, feature selection, creating training data set.

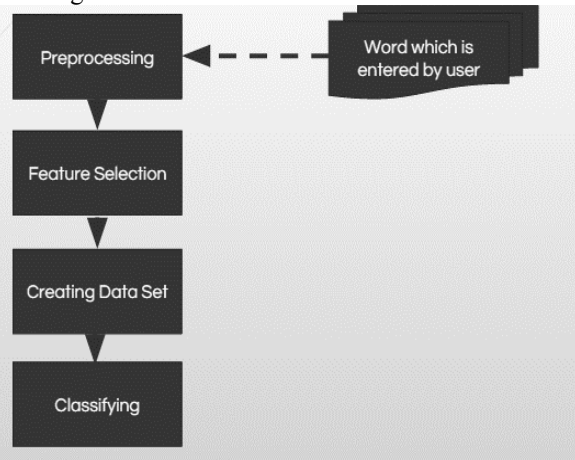

Figure 1. Process of classifying words

In the preprocessing step, the data collected from the application need to be cleaned. There are many data which has the same translation but calculated as different translation just because of case sensitive, or may be some symbol. For example the translation of "Kutub Utara" and "kutub utara", "3G" and "3-G". All the translations are changed into lowercase and be calculated as one translation, then we count and note the number of people who give the same translation for that word (frequency). The number of vote up and vote down are automatically summarized follow the words elimination/merging, as well as calculating the average of the lifetime. Other than that, removing some stop-word and punctuation is done to make the translation cleaner. Recall that the Translator Gator consider about the meaning of the translation, not only looking at the syntax translation, so the translation words should not be filtered by all the stop-word in list because it can cause changes in meaning. About 10,000 translation collected data is processed to be tagged into their class. We generate two classes: correct class and incorrect class. The tagging is done manually, checking by looking at the meaning of the translation.

Next, the data go through the second steps, which is feature selection. The data is analyzed and produced some variables to be considered to find the pattern of correct and incorrect class. Those variables are: 1) Number of vote up and vote down for each word, 2) frequency of a translation word, 3) Lifetime of a word, 4) Weight of vote up and vote down, 5) Differential of weighted vote up and vote down, 6)Class of each word. The weight of vote up is number of vote up / lifetime, and the weight of vote down is number of vote down / lifetime. Meanwhile the weight of differential is weight of vote up - weight of vote down. The last step before getting into training process is creating data set. The data set attributes are the source word (origin word), the translation word, number of vote up, number of vote down, frequency, lifetime, weight of vote up, weight of vote down, differential score, and the class.

3 Comparing Feature Combination

3.1 Feature Lists

There are some feature that would be evaluate in this study to classify translation result to be correct and incorrect translation. Below are the details of the features:

1. Frequency: the number of same 'translation' for one 'origin word'.
2. Vote up: the number of the agree vote for an entry of proposed translation by other user
3. Vote down: the number of the disagree vote for an entry of proposed translation by other user
4. Lifetime: period of an entry of proposed translation in the system

$$lifetime\_origin\_word_i\ (seconds) = time_{now} - appearance\_time_i \quad \text{.....} \quad (1)$$

where:
i: an entry of proposed translation by user
timenow: 23 July 2016

5. Weighted Vote Up: : the ratio between the number of the agree vote for an entry of proposed translation by other user and its lifetime

$$weighted_{vote\_up_i} = \frac{vote\_up_i}{lifetime\_origin\_word_i} \quad \text{.....} \quad (2)$$

6. Weighted Vote Down: the ratio between the number of the agree vote for an entry of proposed translation by other user and its lifetime

$$weighted_{vote\_down_i} = \frac{vote\_down_i}{lifetime\_origin\_word_i} \quad \text{.....} \quad (3)$$

7. Difference Score Between Weighted Vote Up and Vote Down

$$diff_{score} = weighted_{vote\_up_i} - weighted_{vote\_down_i} \quad \text{.....} \quad (4)$$

3.2 Experimental Results

We used Weka (Waikato Environment for Knowledge Analysis) application as a tool. Weka is a software containing collection of machine learning algorithms for data mining tasks developed by Waikato University. Weka contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization.

We use some classification methods in Weka to evaluate the feature lists and its combination that affect the translation results. The methods are Naïve Bayes, Random Forest, SVM, and J-48. The results are shown in the following tables.

No	Feature	Evaluation	Naïve Bayes	Random Forest	SVM	J-48
1	Frequency | Vote up | Vote down | Lifetime |	Correctly Classified Instances	81.37%	81.37%	81.37%	81.37%
		a classified as a (TN)	0	0	0	0
		a classified as b (FN)	1487	1487	1487	1487
		b classified as a (FP)	0	0	0	0
		b classified as b (TP)	6493	6493	6493	6493
2	Vote up, Vote down, Lifetime	Correctly Classified Instances	80.53%	71.32%	81.37%	81.5%
		a classified as a (TN)	99	303	0	38
		a classified as b (FN)	1388	1184	1487	1449
		b classified as a (FP)	166	1105	0	28
		b classified as b (TP)	6327	5388	6493	6465
3	Weighted Vote up, Weighted Vote down	Correctly Classified Instances	81.10%	79.95%	81.37%	81.37%
		a classified as a (TN)	10	66	0	0
		a classified as b (FN)	1477	1421	1487	1487
		b classified as a (FP)	31	179	0	0
		b classified as b (TP)	6462	6314	6493	6493
4	Frequency, Vote up, Vote down	Correctly Classified Instances	47.01%	81.04%	81.37%	81.37%
		a classified as a (TN)	1254	28	0	0
		a classified as b (FN)	233	1459	1487	1487
		b classified as a (FP)	3996	54	0	0
		b classified as b (TP)	2497	6439	6493	6493
5	Frequency, Weigh	Correctly Classif	73.55%	80.06%	81.37%	81.37%

No	Feature	Evaluation	Naïve Bayes	Random Forest	SVM	J-48
	ted Vote up, Weighted Vote down	ied Instances				
		a classified as a (TN)	313	68	0	0
		a classified as b (FN)	1174	1419	1487	1487
		b classified as a (FP)	937	172	0	0
		b classified as b (TP)	5556	6321	6493	6493
6	Diff score	Correctly Classified Instances	81.37%	79.66%	81.37%	81.37%
		a classified as a (TN)	0	57	0	0
		a classified as b (FN)	1487	1430	1487	1487
		b classified as a (FP)	0	193	0	0
		b classified as b (TP)	6493	6300	6493	6493
7	Frequency, Diff score	Correctly Classified Instances	45.11%	80.01%	81.37%	81.37%
		a classified as a (TN)	1252	55	0	0
		a classified as b (FN)	235	1432	1487	1487
		b classified as a (FP)	4145	163	0	0
		b classified as b (TP)	2348	6330	6493	6493

Notes:
a : incorrect translation
b : correct translation

Table 1. Experimental Result of Feature Discovery

4 Conclusion

Based on the experimental results, vote up, vote down and lifetime affect the user translation in crowdsourcing incentive method with the percentage of correctly classified instance 80.53% and 71.32% by Naïve Bayes and Random Forest methods. Weighted vote up and weighted vote down as the ratio of vote up, vote down and lifetime respectively, give a better results with the percentage of correctly classified instance 81.10% and 79.95%. Moreover, the accuracy of combination of frequency, weighted vote up, and weighted vote down features slightly decrease from just weighted vote up and vote down features with the percentage of correctly classified instance 73.55% and 80.06%. However, the number of false positive and false negative are less than before. It can be conclude that frequency feature can be combined with weighted vote up and weighted vote down to classified the correct and incorrect translation.

Diff score as the differentiation of weighted vote up and weighted vote down feature give a slightly worst result. However, the combination of frequency and diff score give a better accuracy with the percentage of correctly classified instance 80.01% based on random forest method. So, frequency and diff score can be used as alternate to classified the correct and incorrect translation.

References

Gesu, V. Di., Isgro, F., Tegolo, D., Trucco, E., (2003). *Finding Essential Features for Tracking Starfish in a Video Sequence. In Proceedings of the 12th International Conference on Image Analysis and Processing*

(ICIAP 03). Mantova, Italy: Institute of Electrical and Electronics Engineers.

Dahl, Anders Lindbjerg., Aanaes, Henrik., Pedersen, Kim Steenstrup. (2011). *Finding the Best Feature Detector-Descriptor Combination*. International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission. Hangzhou, China: Institute of Electrical and Electronics Engineers.

Deng, Da., Simmermacher, Christian, Cranefield, Stephen. (2006). *Finding the Right Features for Instrument Classification of Classical Music*. In Proceedings of the International Workshop on Integrating AI and Data Mining (AIDM 06). Hobart, Tasmania, Australia: Institute of Electrical and Electronics Engineers.

MIDVIS: Pyramid Visualization of SDGs Understanding in Indonesian Community

Retno Kusumaningrum
Universitas Diponegoro
retno_ilkom@undip.ac.id

Iwan Njoto Sandjaja
Universitas Kristen Petra
iwanns@petra.ac.id

Muhammad Rheza
Pulse Lab Jakarta
muhammad.rheza@un.or.id

Hendra Bunyamin
Universitas Kristen Maranatha
hendra.bunyamin@it.maranatha.edu

Ni Made Satvika Iswari
Universitas Multimedia Nusantara
satvika@umn.ac.id

Abstract

Translator Gator is a crowdsourcing translation with an incentive inspired by the need to socialise the 17 Sustainable Development Goals (SDGs) and 10 of Indonesian government's programmes. The huge number of Translator Gator data makes it hard to see and understand the data especially for decision maker users or end users. Therefore, MIDVIS (Pyramid Visualization) is designed and created to solve that problem. We also offer two alternative visualizations: Recurvise Aristotle's Square of Opposition visualization and Zoomable Wordmap visualization. MIDVIS visualizes the inverse rank, which is the most important information in the smallest part on the top of the pyramid. Recurvise Aristotle's Square of Opposition visualizes the most understood word and the most confusing word. Users can see the less understood and less confusing words with a mouse click to zoom in the recursive square. Zoomable Wordmap is useful to compare the level of understanding of each area/language. For further works, a creative essay about SDGs and Indonesia's programme can be constructed automatically by a computer using a computational linguistic method.

1 Introduction

On September 25th 2015, the United Nations adopted the SDGs (Sustainable Development Goals) to end poverty, to fight inequality and injustice, and to tackle climate change by 2030. There are a total of 17 goals that people in all parts of the world are expected to understand and put into reality. According to UNDP reports, it comprises into national, regional, global, and thematic level. National reporting is the most significant level of reporting which is based on Complementary National Indicators that address each country's specific challenges, priorities, and preferences (Mothe, Espey, and Schmidt-Traub, 2015). However, the report only focuses on the pre-defined indicators.

Moreover, we need another approach to see how people from different area in a country understand SDGs. This can be measured by inviting citizen to participate in translating and finding the synonym of each term that represents each SDGs criteria. Hence, Pulse Lab Jakarta created Translator Gator to gather the data translation of English terms which represent SDGs goals into 6 languages, i.e. Indonesian, Sundanese, Buginese, Malay, Minangkabau, and Javanese and also to find their synonym.

However, the huge number of Translator Gator data impact hardly to understand by the level of user, such as decision maker users or end users. Therefore, we propose a novel visualization to have better information about SDGs understanding in Indonesian Community. The visualization is called MIDVIS (Pyramid Visualization). Pyramid shape is expected to visualize the inverse rank, which is the most important information in the smallest part on the top of the pyramid.

2 Related Works

Data visualization is used to give better understanding about information to be delivered to the audience. The use of images to represent information provides a powerful means both to make sense of data and to communicate what we have discovered to others. Rooij, Odijk, and Rijke (2013) in their work visualized the stream of themes discussed in Politics. They described ThemeStreams as a demonstrator that mapped political discussions related to themes and influence makers and illustrated how this mapping was used in an interactive visualization that showed us which themes were being discussed. In the initial usability studies that have been carried out, the main findings indicate that ThemeStreams could be understood intuitively, and inspection of parts of any query was easy to do.

3 Data and Computation

As mentioned in the previous section, we use dataset from Translator Gator, i.e. a game that builds taxonomies for research initiatives. Translator Gator has achieved more than 109,000 user across

Indonesia. The dataset consists of 1609 English terms which are all translated into Indonesian, 460 terms or about 28.59% translated into Sundanese, 91 terms or about 5.66% translated into Buginese, 414 terms or about 25.73% translated into Malay, 184 terms or about 11.44% translated into Minangkabau, and 1339 words or about 83.22% translated into Javanese.

We summarize the dataset into three levels which are subsequently used as data source for each visualization. For the first level, the dataset is summarized on the basis of the SDGs criteria and sorted from the most confusing criteria to the least confusing criteria. For the second level, the output from previous level is grouped by its 6 translated languages and sorted by level of understanding. Furthermore, the whole terms from the second level (for each category and translated language) are grouped into two classes, namely confusing terms and understood terms.

4 Interaction and Visualization

The MIDVIS homepage displays a pyramid consisting of 17 SDGs sorted from the most confusing criteria to the least confusing criteria. Users can click an SDG criteria and the MIDVIS will display a map of Indonesia. The colors of Indonesia map area is gradated from the chosen color representation of SDG toward lighter color in 6 level gradation. People in regions with the darker color are the ones who understand SDG better while people in regions with the lighter color have the least understanding of SDG. Users can see some information, such as location name, two languages used by local people in that location, the total number of all terms that have been translated, and some example terms that have been translated while hovering at some area in the map.

The last visualization is a seesaw model, which represents two groups of terms, i.e. confusing and understood terms. Initially, the seesaw is balanced,

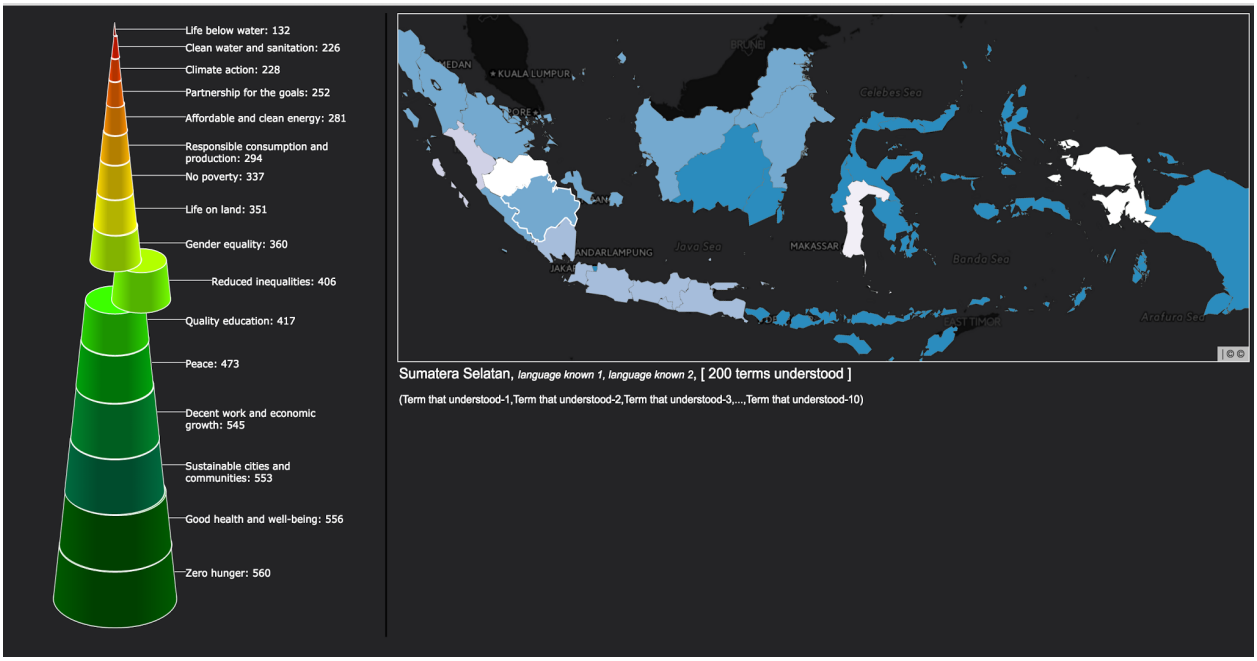

Figure 1. MIDVIS

Figure 2. Seesaw Visualization

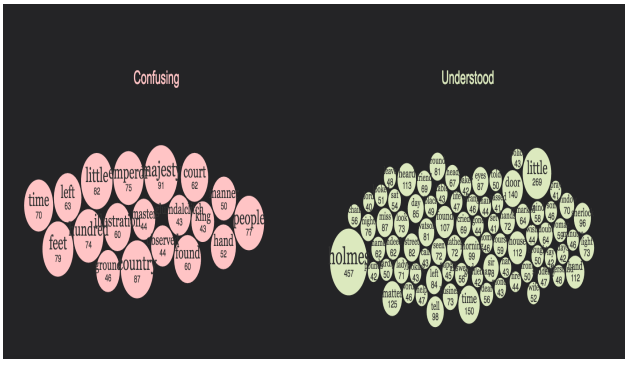

Figure 3. Alternative representation for words using Bubble Visualization

with the left side representing a group of confusing terms and the right side representing a group of understood terms. The terms in each group are represented as bubbles. The size of the bubble represents the weight of the word. Subsequently, the seesaw will animate and the final position is determined by the total weight of terms in each group. The group with higher number of terms has more weight than the other group.

5 Alternatives Visualization

The first alternative visualization for replacing a seesaw model is using recursive square of opposition. The idea comes from Aristotle's square of opposition. Aristotle used the diagonal to represent contradictory terms. By using the opposition square recursively, we can represent many terms infinitely. The main advantage of this visualization is the user will see directly the most confusing terms and most understood terms. Imagine that the four most biggest bubbles from both ends of seesaw will be put in the outer opposition square. The next four from both ends of seesaw will be put in the second square which is smaller and rotated. By doing this we can put all the terms into the square of opposition. The users will always see the biggest bubbles first. If they want to see the smaller bubbles, they have to click to zoom in and the small square will grow larger replacing the outer square. This new and original visualization will help the users to see the most important terms (the biggest bubble) first and require more effort to see the less important terms. Widdows (2004) showed a similar visualization but without a recursive part.

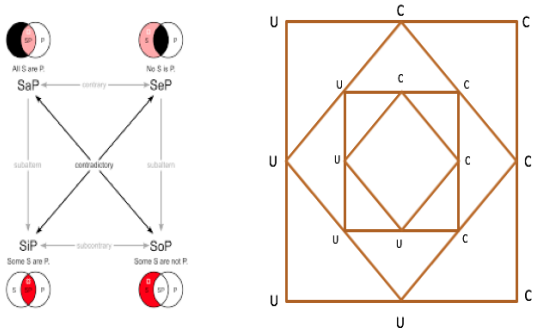

Figure 4. Recursive square of opposition visualization

The second alternative is Zoomable Wordmap. In this visualization, all terms in English will be sorted alphabetically. Each term will be represented by a small color coded square. The color will be the indicator how the users understand the word. If a user has high understanding of the word, the color will be red. Otherwise, the color will be blue. The gradation from blue to red will indicate the level of the user's understanding. The user can compare the same wordmap for different languages/areas because the position of the word is fixed. If the user needs to see what word is represented by a square, the user can click to zoom in and see the word.

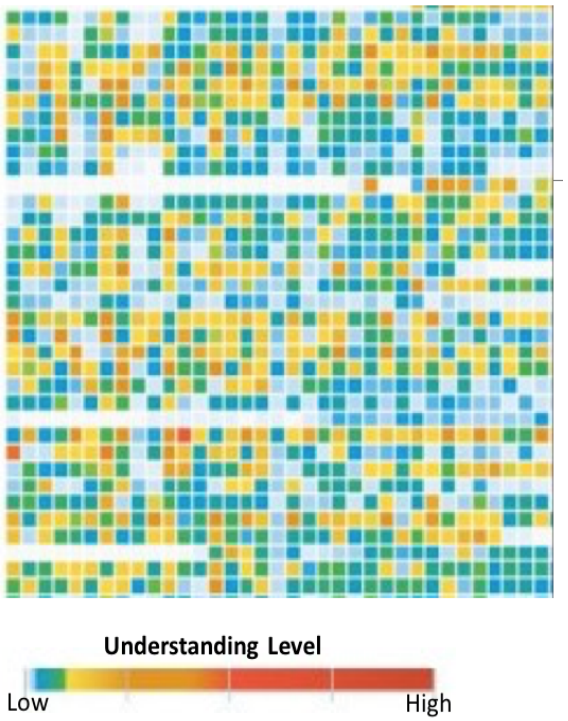

Figure 6. Zoomable Wordmap Visualization

6 Conclusion

We have demonstrated a new way of visualizing SDGs indicator. This visualization will enable the decison maker to act promptly in order to improve communication between Indonesian government and community. For the next step, we can use the same data to creatively compose an essay explaining 17 SDGs in Indonesian informal languages by applying a computational linguistic method.

References

Dominic Widdows. 2004. *Geometry and Meaning*, Lecture Notes (Book 172). CSLI Publication, Stanford University, CA.

Ork de Rooij, Daan Odijk, and Maarten de Rijke. 2013. *ThemeStreams: Visualizing The Stream Of Themes Discussed In Politics.* SIGIR'13: 36th international ACM SIGIR conference on Research and development in information retrieval.

Eve de la Mothe, Jessica Espey and Guido Schmidt-Traub. (2016, August 5). *Measuring Progress on the SDGs: Multi-level Reporting* [Report GSDR 2015 Brief]. Retrieved from https://sustainabledevelopment.un.org/content/documents/6464102-Measuring%20Progress%20on%20the%20SDGs%20%20%20Multi-level%20Reporting.pdf

Pulse Lab Jakarta is grateful for the generous support from
the Department of Foreign Affairs and Trade of the Government of Australia.