Generating Similarity Cluster of Indonesian Languages with Semi-Supervised Clustering

by Nasution Arbi Haza

Submission date: 30-Oct-2019 04:04PM (UTC+0800)
Submission ID: 1203421018
File name: 3.pdf (2.33M)
Word count: 5219
Character count: 26656

International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 1, February 2019, pp. 531∼ 538
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i1.pp531-538

Generating similarity cluster of Indonesian languages with semi-supervised clustering

Arbi Haza Nasution[1,3], Yohei Murakami[2], and Toru Ishida[3]
[1,3]Department of Social Informatics, Kyoto University, Japan
[2]College of Information Science and Engineering, Ritsumeikan University, Japan
[1]Department of Informatics Engineering, Universitas Islam Riau, Indonesia

Article Info

Article history:
Received Jan 11, 2018
Revised Jul 6, 2018
Accepted Aug 22, 2018

Keywords:
Lexicostatistic
Language Similarity
Hierarchical Clustering
K-means Clustering
Semi-Supervised Clustering

ABSTRACT

Lexicostatistic and language similarity clusters are useful for computational linguistic researches that depends on language similarity or cognate recognition. Nevertheless, there are no published lexicostatistic/language similarity cluster of Indonesian ethnic languages available. We formulate an approach of creating language similarity clusters by utilizing ASJP database to generate the language similarity matrix, then generate the hierarchical clusters with complete linkage and mean linkage clustering, and further extract two stable clusters with high language similarities. We introduced an extended k-means clustering semi-supervised learning to evaluate the stability level of the hierarchical stable clusters being grouped together despite of changing the number of cluster. The higher the number of the trial, the more likely we can distinctly find the two hierarchical stable clusters in the generated k-clusters. However, for all five experiments, the stability level of the two hierarchical stable clusters is the highest on 5 clusters. Therefore, we take the 5 clusters as the best clusters of Indonesian ethnic languages. Finally, we plot the generated 5 clusters to a geographical map.

Corresponding Author:
Arbi Haza Nasution,
Department of Social Informatics, Kyoto University, Japan,
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan.
+818078554376
Email: arbi@ai.soc.i.kyoto-u.ac.jp, arbi@eng.uir.ac.id

1. INTRODUCTION

Nowadays, machine-readable bilingual dictionaries are being utilized in actual services [1] to support intercultural collaboration [2, 3, 4] and other research domains [5, 6, 7, 8, 9], but low-resource languages lack such sources. Indonesia has a population of 221,398,286 and 707 living languages which cover 57.8% of Austronesian Family and 30.7% of languages in Asia [10]. There are 341 Indonesian ethnic languages facing various degree of language endangerment (trouble / dying) where some of the native speaker do not speak Bahasa Indonesia well since they are in remote areas. Unfortunately, there are 13 Indonesian ethnic languages which already extinct. In order to save low-resource languages like Indonesian ethnic languages from language endangerment, prior works tried to enrich the basic language resource, i.e., bilingual dictionary [11, 12, 13, 14]. Those previous researchers require lexicostatistic/language similarity clusters of the low-resource languages to select the target languages. However, to the best of our knowledge, there are no published lexicostatistic/language similarity clusters of Indonesian ethnic languages. To fill the void, we address this research goal: Formulating an approach of creating a language similarity cluster. We first obtain 40-item word lists from the Automated Similarity Judgment Program (ASJP), further generate the language similarity matrix, then generate the hierarchical and k-means clusters, and finally plot the generated clusters to a map.

2. AUTOMATED SIMILARITY JUDGMENT PROGRAM

Historical linguistics is the scientific study of language change over time in term of sound, analogical, lexical, morphological, syntactic, and semantic information [15]. Comparative linguistics is a branch of historical linguistics that is concerned with language comparison to determine historical relatedness and to construct language families [16]. Many methods, techniques, and procedures have been utilized in investigating the potential distant genetic relationship of languages, including lexical comparison, sound correspondences, grammatical evidence, borrowing, semantic constraints, chance similarities, sound-meaning isomorphism, etc [17]. The genetic relationship of languages is used to classify languages into language families. Closely-related languages are those that came from the same origin or proto-language, and belong to the same language family.

Swadesh List is a classic compilation of basic concepts for the purposes of historical-comparative linguistics. It is used in lexicostatistics (quantitative comparison of lexical cognates) and glottochronology (chronological relationship between languages). There are various version of swadesh list with a number of words equal 225 [18], 215 & 200 [19], and lastly 100 [20]. To find the best size of the list, Swadesh states that "The only solution appears to be a drastic weeding out of the list, in the realization that quality is at least as important as quantity. Even the new list has defects, but they are relatively mild and few in number." [21]

A widely-used notion of string/lexical similarity is the edit distance or also known as Levenshtein Distance (LD): the minimum number of insertions, deletions, and substitutions required to transform one string into the other [22]. For example, LD between "kitten" and "sitting" is 3 since there are three transformations needed: kitten sitten (substitution of "s" for "k"), sitten sittin (substitution of "i" for "e"), and finally sittin sitting (insertion of "g" at the end).

There are a lot of previous works using Levenshtein Distances such as dialect groupings of Irish Gaelic [23] where they gather the data from questionnaire given to native speakers of Irish Gaelic in 86 sites. They obtain 312 different Gaelic words or phrases. Another work is about dialect pronunciation differences of 360 Dutch dialects [24] which obtain 125 words from Reeks Nederlandse Dialectatlassen. They normalize LD by dividing it by the length of the longer alignment. [25] measure linguistic similarity and intelligibility of 15 Chinese dialects and obtain 764 common syllabic units. [26] define lexical distance between two words as the LD normalized by the number of characters of the longer of the two. [27] extend Petroni definition as LDND and use it in Automated Similarity Judgment Program (ASJP).

The ASJP, an open source software was proposed by [28] with the main goal of developing a database of Swadesh lists [21] for all of the world's languages from which lexical similarity or lexical distance matrix between languages can be obtained by comparing the word lists. The classification is based on 100-item reference list of Swadesh [20] and further reduced to 40 most stable items [29]. The item stability is a degree to which words for an item are retained over time and not replaced by another lexical item from the language itself or a borrowed element. Words resistant to replacement are more stable. Stable items have a greater tendency to yield cognates (words that have a common etymological origin) within groups of closely related languages.

3. LANGUAGE SIMILARITY CLUSTERING APPROACH

We formalize an approach to create language similarity clusters by utilizing ASJP database to generate the language similarity matrix, then generate the hierarchical clusters, and further extract the stable clusters with high language similarities. The hierarchical stable clusters are evaluated utilizing our extended k-means clustering. Finally, the obtained k-means clusters are plotted to a geographical map. The flowchart of the whole process is shown in Figure 1.

In this paper, we focus on Indonesian ethnic languages. We obtain words list of 119 Indonesian ethnic languages with the number of speakers at least 100,000. However, it is difficult to classify 119 languages and obtain a valuable information from the generated clusters, therefore, we further filtered the target languages based on the number of speaker and availability of the language information in Wikipedia. We obtain 32 target languages as shown in Table 1 from the intersection between 46 Indonesian ethnic languages with number of speaker above 300,000 provided by Wikipedia and 119 Indonesian ethnic languages with number of speaker above 100,000 provided by ASJP.

We further generate the similarity matrix of those 32 languages as shown in Figure 2. We added a white-red color scale where white color means the two languages are totally different (0% similarity) and the reddest color means the two languages are exactly the same (100% similarity). For a better clarity and to avoid redundancy, we only show the bottom-left part of the table. The headers follow the language code in Table 1.

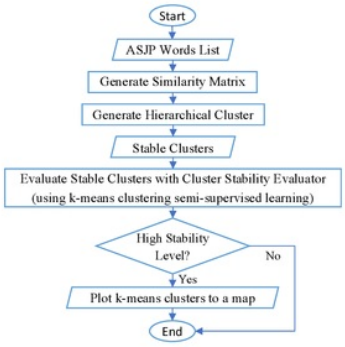


Figure 1. Flowchart of Generating Language Similarity Clusters

Table 1. List of 32 Indonesian Ethnic Languages Ranked by Population According to ASJP database

Code	Population	Language	Code	Population	Language
L 1	232004800	INDONESIAN	L 17	1000000	GORONTALO
L 2	84300000	OLD OR MIDDLE JAVANESE	L 18	1000000	JAMBI MALAY
L 3	34000000	SUNDANESE	L 19	900000	MANGGARAI
L 4	15848500	MALAY	L 20	770000	NIAS NORTHERN
L 5	15848500	PALEMBANG MALAY	L 21	750000	BATAK ANGKOLA
L 6	6770900	MADURESE	L 22	700000	UAB METO
L 7	5530000	MINANGKABAU	L 23	600000	KARO BATAK
L 8	5000000	BUGINESE	L 24	500000	BIMA
L 9	5000000	BETAWI	L 25	470000	KOMERING
L 10	3502300	BANJARESE MALAY	L 26	350000	REJANG
L 11	3500032	ACEH	L 27	331000	TOLAKI
L 12	3330000	BALI	L 28	300000	GAYO
L 13	2130000	MAKASAR	L 29	300000	MUNA
L 14	2100000	SASAK	L 30	250000	TAE
L 15	2000000	TOBA BATAK	L 31	245020	AMBONESE MALAY
L 16	1100000	BATAK MANDAILING	L 32	230000	MONGONDOW

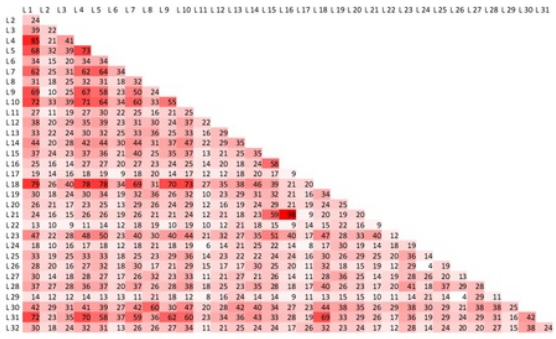
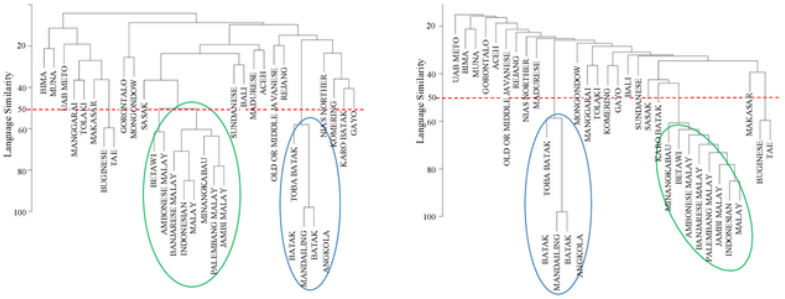


Figure 2. Lexicostatistic / Similarity Matrix of 32 Indonesian Ethnic Languages by ASJP (%)

Hierarchical clustering is an approach which builds a hierarchy from the bottom-up, and does not require us to specify the number of clusters beforehand. The algorithm works as follows: (1) Put each data point in its own cluster; (2) Identify the closest two clusters and combine them into one cluster; (3) Repeat the above step until all the data points are in a single cluster. Once this is done, it is usually represented by a dendrogram like structure. There are a few ways to determine how close two clusters are: (1) Complete linkage clustering: find the maximum possible distance between points belonging to two different clusters; (2) Single linkage clustering: find the minimum possible distance between points belonging to two different clusters; (3) Mean/Average

Linkage clustering: find all possible pairwise distances for points belonging to two different clusters and then calculate the average; (4) Centroid linkage clustering: find the centroid of each cluster and calculate the distance between centroids of two clusters. Complete linkage and mean (average) linkage clustering are the ones used most often. We generate the distance matrix from the similarity matrix shown in Figure 2 and further generate the hierarchical clusters with hclust function with a complete linkage clustering method as shown in Figure 3(a) and a mean linkage clustering method as shown in Figure 3(b) using R, a free software environment for statistical computing and graphics.



(a) Method: Complete (b) Method: Average

Figure 3. Hierarchical Clusters Dendogram of 32 Indonesian Ethnic Languages.

From those two hierarchical clusters in Figure 3, we select two stable clusters that always grouped together despite of changing the linkage clustering method. The first cluster consists of TOBA BATAK, BATAK MANDAILING, and BATAK ANGKOLA, while the second cluster consists of MINANGKABAU, BETAWI, AMBONESE MALAY, BANJARESE MALAY, PALEMBANG MALAY, JAMBI MALAY, MALAY, and Indonesia. Since the two stable custers have language similarities above 50% between the languages, they are good clusters to be referred when selecting target languages for computational linguistic researches that depends on language similarity or cognate recognition for inducing bilingual lexicons from the target languages [11, 12, 14, 30]. The two clusters are actually enough for selecting the target languages for those researches. However, we still need to evaluate the stability of those clusters and we also need to identify the low language similarities clusters in order to grasp the whole picture of Indonesian ethnic languages. Thus, we utilize the alternative clustering approach which is a k-means clustering.

K-means clustering is an unsupervised learning algorithm that tries to cluster data based on their similarity. Unsupervised learning means that there is no outcome to be predicted, and the algorithm just tries to find patterns in the data. In k-means clustering, we have to specify the number of clusters we want the data to be grouped into. The algorithm works as follows: (1) The algorithm randomly assigns each observation to a cluster, and finds the centroid of each cluster; (2) Then, the algorithm iterates through two steps: (2a) Reassign data points to the cluster whose centroid is closest; (2b) Calculate new centroid of each cluster. These two steps are repeated until the within cluster variation cannot be reduced any further. The within cluster variation is calculated as the sum of the euclidean distance between the data points and their respective cluster centroids.

It is well known that standard agglomerative hierarchical clustering techniques are not tolerant to noise [31, 32]. There are many previous works on finding clusters which robust to noise [33, 34, 35]. However, to evaluate the stability of the hierarchical stable clusters, we introduced a simple approach of calculating their stability level of being grouped together despite of changing the number of k-means clusters. We extend the k-means clustering unsupervised learning to a k-means clustering semi-supervised learning as shown in Algorithm 1 by labeling the two hierarchical stable clusters beforehand.

Algorithm 1: Cluster Stability Evaluator

Input: $similarityMatrix$, $stableClusters$, $minimumK$, $maximumTrial$;
Output: $stabilityLevel$

$trial \leftarrow 1$;
$currentK \leftarrow minimumK$;
$maximumK \leftarrow length(similarityMatrix)$;
$scale2D \leftarrow cmdscale(similarityMatrix)$; // multidimensional to 2D scaling
while $currentK <= maximumK$ **do**
  $successfulTrial \leftarrow 0$; // initialized for each $currentK$
  while $trial <= maximumTrial$ **do**
    $kClusters \leftarrow kmeans(scale2D, currentK)$;
    if $stableClusters$ *distinctly found in* $kClusters$ **then**
      $successfulTrial++$;
      $trial++$; // try again with the same number of cluster ($currentK$)
    end
  end
  $stabilityLevel[currentK] \leftarrow successfulTrial/maximumTrial$;
  $currentK++$; // increase the number of clusters
  $trial \leftarrow 1$ // reset the number of trial
end
return $stabilityLevel$;

4. RESULT AND DISCUSSION

Initially, we manually conduct several trials to estimate the minimum and maximum number of k-means cluster to obtain clusters which consist of the stable clusters distinctly. Based on the initial trials, we estimate the $minimum_k = 4$ and $maximum_k = 21$. Then, we calculate the stability level of the two hierarchical stable clusters where the number of clusters ranging from $minimum_k = 4$ to $maximum_k = 21$ following Algorithm 1. We have five sets of experiments with the $maximum_trial$ equals 50, 500, 5,000, 50,000, and 500,000. In each experiment, a stability level of the two hierarchical stable clusters is measured for each number of k-means clusters by calculating the success rate of obtaining the two hierarchical stable clusters in the generated k-clusters as shown in Figure 4.

The higher the number of the trial, the more likely we can distinctly find the two hierarchical stable clusters in the generated k-clusters with a big number of clusters. For example, within 50 trials, we can not find the two hierarchical stable clusters distinctly in the generated k-clusters for big number of clusters ($k > 14$). However, within 50,000 and 500,000 trials, we can find the two hierarchical stable clusters distinctly in the generated k-clusters for all number of clusters between the $minimum_k = 4$ and the $maximum_k = 21$, even though the success rate is getting lower as the number of clusters increases. For all five experiments, the stability level of the two hierarchical stable clusters is the highest (0.78) on 5 clusters.

Therefore, we take the 5 clusters as shown in Figure 5 as the best clusters of Indonesian ethnic languages to be referred when selecting target languages for computational linguistic researches that depends on language similarity or cognate recognition. We further plot the 5 clusters to a geographical map as shown in Figure 6.

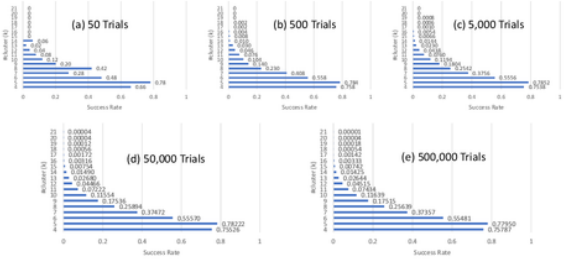


Figure 4. Obtaining Stable Clusters in n Trials

Figure 5. K-means Clusters of 32 Indonesian Ethnic Languages – 5 Clusters

Figure 6. Similarity Clusters Map of 32 Indonesian Ethnic Languages – 5 Clusters

5. CONCLUSION

We utilized ASJP database to generate the language similarity matrix, then generate the hierarchical clusters with complete linkage and mean linkage clustering, and further extract two stable clusters with the highest language similarities. We apply our extended k-means clustering semi-supervised learning to evaluate the stability level of the hierarchical stable clusters being grouped together despite of changing the number of clusters. The higher the number of the trial, the more likely we can distinctly find the two hierarchical stable clusters in the generated k-clusters. However, for all five experiments, the stability level of the two hierarchical stable clusters is the highest (0.78) on 5 clusters. Therefore, we take the 5 clusters as the best clusters of Indonesian ethnic languages to be referred to select target languages for computational linguistic researches that depends on language similarity or cognate recognition. Finally, we plot the generated 5 clusters to a geographical map. Our algorithm can be used to find and evaluate other stable clusters of Indonesian ethnic languages or other language sets.

ACKNOWLEDGEMENT

This research was partially supported by a Grant-in-Aid for Scientific Research (A) (17H00759, 2017-2020) and a Grant-in-Aid for Young Scientists (A) (17H04706, 2017-2020) from Japan Society for the Promotion of Science (JSPS). The first author was supported by Indonesia Endownment Fund for Education (LPDP).

REFERENCES

[1] T. Ishida,Y. Murakami, D. Lin, T. Nakaguchi and M. Otani, "Language Service Infrastructure on the Web: The Language Grid," *IEEE Computer*, vol. 51, Issue 6, pp. 72-81, June, 2018.

[2] T. Ishida, "Intercultural collaboration and support systems: A brief history," in *International Conference on Principles and Practice of Multi-Agent Systems (PRIMA 2016)*, pages 3-19. Springer, 2016.

[3] A. H. Nasution, N. Syafitri, P. R. Setiawan and D. Suryani, "Pivot-Based Hybrid Machine Translation to Support Multilingual Communication," in *International Conference on Culture and Computing (Culture and Computing), Kyoto, Japan*, 2017, pp. 147-148. doi: 10.1109/Culture.and.Computing.2017.22.
[4] A. H. Nasution, "Pivot-Based Hybrid Machine Translation to Support Multilingual Communication for Closely Related Languages," *World Transactions on Engineering and Technology Education*, 16, 2, 12-17, 2018.
[5] A. H. Nasution, Y. Murakami, and T. Ishida. "Designing a Collaborative Process to Create Bilingual Dictionaries of Indonesian Ethnic Languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Paris, France, 3397-3404, 2018.
[6] R. Tanaka, Y. Murakami and T. Ishida, "Context-Based Approach for Pivot Translation Services," in *International Joint Conference on Artificial Intelligence (IJCAI-09)*, pp.1555-1561, 2009.
[7] E. W. Pamungkas, R. Sarno, and A. Munif, "B-BabelNet: Business-Specific Lexical Database for Improving Semantic Analysis of Business Process Models," *Telkomnika*, 15(1), 407, 2017.
[8] H. Hassan, "A framework for Arabic concept-level sentiment analysis using SenticNet," *International Journal of Electrical and Computer Engineering (IJECE)*, 8(6), 2018.
[9] P. Bajpai, P. Verma and S. Q. Abbas, "Two Level Disambiguation Model for Query Translation," *International Journal of Electrical and Computer Engineering (IJECE)*, 8(5), 2018.
[10] Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.), *Ethnologue: Languages of the World, Eighteenth edition. Dallas, Texas: SIL International.* Online version: http://www.ethnologue.com, 2015.
[11] A. H. Nasution, Y. Murakami, and T. Ishida, "Constraint-based bilingual lexicon induction for closely related languages," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 3291-3298, Paris, France, May, 2016.
[12] A. H. Nasution, Y. Murakami and T. Ishida, "A generalized constraint approach to bilingual dictionary induction for low-resource language families," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17, 2, Article 9 (November 2017), 29 pages, 2017.
[13] A. H. Nasution, Y. Murakami and T. Ishida, "Plan Optimization for Creating Bilingual Dictionaries of Low-Resource Languages," in *International Conference on Culture and Computing (Culture and Computing), Kyoto, Japan*, 2017, pp. 35-41. doi: 10.1109/Culture.and.Computing.2017.21.
[14] M. Wushouer, D. Lin, T. Ishida and K. Hirayama, "A constraint approach to pivot-based bilingual dictionary induction," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 15(1):4:1-4:26, November, 2015.
[15] L. Campbell. *Historical Linguistics.* Edinburgh University Press, 2013.
[16] W. P. Lehmann. *Historical linguistics: an introduction.* Routledge, 2013.
[17] L. Campbell and W.J. Poser. *Language classification. History and method.* Cambridge, 2008.
[18] M. Swadesh, "Salish Internal Relationships," *International Journal of American Linguistics*, vol. 16, 157-167, 1950.
[19] M. Swadesh, "Lexicostatistic Dating of Prehistoric Ethnic Contacts," in *Proceedings of the American Philosophical Society*, vol. 96, 452-463, 1952.
[20] M. Swadesh. *The Origin and Diversification of Language*, Ed. post mortem by Joel Sherzer. Chicago: Aldine, p. 283, 1971.
[21] M. Swadesh, "Towards Greater Accuracy in Lexicostatistic Dating," *International Journal of American Linguistics*, vol. 21, 121-137, 1955.
[22] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet physics doklady*, vol. 10, No. 8, pp. 707-710, 1966.
[23] B. Kessler, "Computational dialectology in Irish Gaelic," in *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics (EACL '95)*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 60-66. DOI: https://doi.org/10.3115/976973.976983
[24] W. J. Heeringa. *Measuring dialect pronunciation differences using Levenshtein distance,* Doctoral dissertation, University Library Groningen, 2004.
[25] C. Tang and V. J. van Heuven, "Predicting mutual intelligibility of Chinese dialects from multiple objective linguistic distance measures," *Linguistics*, 53(2), 285-312, 2015.
[26] F. Petroni and M. Serva, "Language distance and tree reconstruction," *Journal of Statistical Mechanics: Theory and Experiment 2008*, no. 08 (2008): P08012.
[27] S. Wichmann, E. W. Holman, D. Bakker, and C. H. Brown, "Evaluating linguistic distance measures,"

Physica A: Statistical Mechanics and its Applications, 389(17), 3632-3639, 2010.
[28] E.W. Holman, C.H. Brown, S. Wichmann, A. Mu ller, V. Velupillai, H. Hammarstro m, S. Sauppe, H. Jung, D. Bakker and P. Brown, "Automated dating of the world's language families based on lexical similarity," *Current Anthropology 52, 6*, 841-875, 2011.
[29] E. W. Holman, S. Wichmann, C. H. Brown, V. Velupillai, A. Müller, and D. Bakker, "Explorations in automated language classification," *Folia Linguistica*, 42(3-4), 331-354, 2008.
[30] G. S. Mann and D. Yarowsky, "Multipath translation lexicon induction via bridge languages," in *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, Association for Computational Linguistics, 1-8, 2001.
[31] G. Nagy, "State of the art in pattern recognition," in *Proceedings of the IEEE*, 56, no. 5 pp.836-863, 1968.
[32] M. Narasimhan, N. Jojic and J. Bilmes, "Q-clustering," *Advances in Neural Information Processing Systems*, 2006.
[33] M. F. Balcan, Y. Liang and P. Gupta, "Robust hierarchical clustering," *The Journal of Machine Learning Research*, 15(1), 3831-3871, 2014.
[34] S. Guha, R. Rastogi and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes," *Information systems*, 25(5), 345-366, 2000.
[35] P. Langfelder and S. Horvath, "Fast R functions for robust correlations and hierarchical clustering," *Journal of statistical software*, 46(11), 2012.

BIOGRAPHY OF AUTHORS

Arbi Haza Nasution is currently working toward the Ph.D. degree in Social Informatics at Graduate School of Informatics, Kyoto University. He obtained Bachelor Degree in Computer Science from National University of Malaysia in 2010 and obtained Master Degree in Management Information System from National University of Malaysia in 2012. He has been a Lecturer with the Department of Informatics Engineering, Universitas Islam Riau, Indonesia, since 2013. His current research interests include computational linguistics, natural language processing and machine learning.

Yohei Murakami received the Ph.D. degree in informatics from Kyoto University, Kyoto, Japan, in 2006. He has been an Associate Professor with Ritsumeikan University, since 2018. He currently leads the research and development of the Language Grid, the purpose of which is to share various language resources as Web services and enable users to create new services. He received the Achievement Award of the Institute of Electronics, Information and Communication Engineers for this work in 2013. His current research interests include services computing and multiagent systems. He founded the Technical Committee on Services Computing with the Institute of Electronics, Information and Communication Engineers in 2012.

Toru Ishida has been a Professor with Kyoto University, Kyoto, Japan, since 1993. His current research interests include autonomous agents and multiagent systems. He has performed research in the above areas for over 20 years. Since 2006, he has been running the Language Grid Project. Prof. Ishida served as the Program Co-Chair of the second ICMAS, the Chair of the first PRIMA, and the General Co-Chair of the first AAMAS. He was also an Editor-in-Chief of the Journal on Web Semantics (Elsevier) and an Associate Editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence and the Journal on Autonomous Agents and Multi-Agent Systems (Springer). He was a Board Member of the International Foundation on Autonomous Agent and Multiagent Systems. He has also started workshops/conferences on digital cities and intercultural collaboration.

Generating Similarity Cluster of Indonesian Languages with Semi-Supervised Clustering

ORIGINALITY REPORT

16%	18%	14%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

7%

★ datascienceplus.com

Internet Source

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		